JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Protein−Ligand-Based Pharmacophores: Generation and Utility Assessment in Computational Ligand Profiling

Jamel Meslamani,[†] Jiabo Li,[‡] Jon Sutter,[‡] Adrian Stevens,[§] Hugues-Olivier Bertrand,[∥] and Didier Rognan[†,*]

[†]Laboratoire d'Innovation Thérapeutique, UMR7200 Université de Strasbourg/CNRS, 74 route du Rhin, 67400 Illkirch, France

[‡]Accelrys, Inc., 10188 Telesis Court, Suite 100, San Diego, California 92121, United States

[§]Accelrys Ltd., 334 Cambridge Science Park, Cambridge CB4 OWN, England

[∥]Accelrys SARL, Parc Club Orsay Université, 20 Rue Jean Rostand, 91898 Orsay Cédex, France

Supporting Information

Ligand	3B9	55V	AD3	AZZ	BAU	BCZ	CEI	CEL	CT5	DES	ET	GEO	GNT	I84	IMN	IXM	NGH
2D-Sim	40	1	10	3	1	1	7	7	1	1	457	34	1	151	6	1	1
3-D Sim	267	5	6	5	2	4	1	5	1	2	2	4	1	47	6	4	7
Pharm	17	7	4	40	1	3	13	4	11	2	190	34	58	103	17	2	5
Docking	27	32	6	27	6	1	1	5	6	2	118	8	1	309	49	26	8

ABSTRACT: Ligand profiling is an emerging computational method for predicting the most likely targets of a bioactive compound and therefore anticipating adverse reactions, side effects and drug repurposing. A few encouraging successes have already been reported using ligand 2-D similarity searches and protein−ligand docking. The current study describes the use of receptor−ligand-derived pharmacophore searches as a tool to link ligands to putative targets. A database of 68,056 pharmacophores was first derived from 8,166 high-resolution protein−ligand complexes. In order to limit the number of queries, a maximum of 10 pharmacophores was generated for each complex according to their predicted selectivity. Pharmacophore search was compared to ligand-centric (2-D and 3-D similarity searches) and docking methods in profiling a set of 157 diverse ligands against a panel of 2,556 unique targets of known X-ray structure. As expected, ligand-based methods outperformed, in most of the cases, structure-based approaches in ranking the true targets among the top 1% scoring entries. However, we could identify ligands for which only a single method was successful. Receptor−ligand-based pharmacophore search is notably a fast and reliable alternative to docking when few ligand information is available for some targets. Overall, the present study suggests that a workflow using the best profiling method according to the protein−ligand context is the best strategy to follow. We notably present concrete guidelines for selecting the optimal computational method according to simple ligand and binding site properties.

INTRODUCTION

Knowledge on protein−ligand binding data (affinity, structure) is increasing at an amazing pace thanks to public initiatives to homogenize data archival and mining.[1,2] On the target side, the Protein Data Bank[3] stores 78,000 three-dimensional (3-D) structures of proteins and protein−ligand complexes out of which about 10,000 relate to druggable proteins and their ligands.[4] On the ligand side, ChEMBL[5] is a repository of more than five million bioactivity data gathered from literature and addressing one million ligands and 8,700 molecular targets. Computational chemists have rapidly developed so-called chemogenomic methods[6] to mine this vast matrix of experimental data in order to predict novel interactions. Whereas many virtual screening methods[7] (similarity search, pharmacophore mapping, protein−ligand docking) have proven useful to predict novel ligands for a single target, profiling a single ligand against a set of heterogeneous targets has long been neglected. Scientific and economic pressure to design drugs with controlled selectivity profiles[8,9] as well as the recent boost of drug repurposing,[10] led to the development of in silico ligand profiling methods[11] aimed at (i) predicting potential targets (and thus a mechanism of action) for orphan bioactive ligands,[12] (ii) identifying off-targets responsible for side effects and adverse reactions,[13] and (iii) proposing novel targets for existing drugs.[14]

From a conceptual point of view, three groups of methods can be used to predict novel protein−ligand interactions. At the simplest level of theory is the concept that *similar ligands bind to similar targets*. Estimating the similarity between a ligand of interest and target-annotated compounds is thus an easy way to predict novel target-ligand associations[13] and even within

Received: February 10, 2012
Published: April 5, 2012

certain limits binding affinities.[15] Interestingly, 2-D similarity methods have recently been shown to be effective for identifying main targets, whereas 3-D similarity methods were better suited for proposing off-targets.[16] Ligand-centric profiling methods are however restricted to targets for which sufficient ligand information is available. For example, the Similarity Ensemble Approach (SEA) developed by Keiser et al. only applies to 246 targets annotated by more than 100 ligands.[17,18]

A second group of methods relies on the concept that *similar ligands bind to similar binding sites*. Binding site similarity either at the sequence[19] or at the structure level[20] can thus be used as a means to pair an existing ligand (of known binding site) to a novel target as successfully evidenced by several independent reports.[21−23] Again, the method has inherited limitations as it is restrained to the few targets for which a 3-D structure is available (structure-based approach) or to a target subfamily in order to avoid binding site-based misalignments (sequence-based approach).

At the highest level of theory is the last group of approaches focusing on protein−ligand complexes that can be described either as simple 1-D fingerprints,[24] protein−ligand-derived pharmacophores,[25] or protein−ligand docking poses.[26] Identification of novel targets, accounting for main or secondary effects, has been reported in numerous reverse docking studies[12,13,26−28] despite notorious deficiencies of empirical scoring functions to rank order target-ligand complexes by increasing binding free energies.[29,30] Chemogenomic (or proteochemometric) approaches correctly predicting novel binary associations also begin to appear in the literature.[31] Surprisingly, pharmacophores have been widely used in many areas of computer-aided drug design[32] but rarely in target fishing applications. The idea to screen protein−ligand-derived pharmacophores in order to identify potential targets of bioactive ligands was applied by Langer et al. in a series of retrospective screening experiments focusing on small protein−ligand matrices.[33−35] A noticeable hurdle to pharmacophore-based ligand profiling is the automation of relevant pharmacophore perception and generation protocols, mainly due to the difficulty to correctly assign atom types and bond orders from raw PDB files.[25] Up to now, only two pharmacophore collections are available. The Inte:Ligand's Pharmacophore Database (http://www.inteligand.com/pharmdb/) is a private-owned repertoire of 2,500 manually assembled pharmacophore models covering 300 clinically relevant pharmacological targets. The PharmTargetDB includes over 7,000 receptor-based pharmacophore models from 1,500 protein−ligand structures and can be screened via the PharmMapper server (http://59.78.96.61/pharmmapper/).[36] Only the Inte:Ligand collection has been successfully screened to identify novel targets (acetylcholinesterase, human rhinovirus coat protein, and cannabinoid receptor type 2) for secondary metabolites from the medicinal plant *Ruta graveolens*.[37]

There are two objective reasons explaining why pharmacophore-based target identification has not become yet a standard in silico ligand profiling method: (i) the abscence of an exhaustive collection of protein−ligand based and(or) ligand-based pharmacophore databases, and (ii) the lack of clear benchmarks comparing the later approach to commonly used strategies (2-D and 3-D ligand similarity search, protein−ligand docking) in computational profiling. We herewith present PharmaDB, the largest ever reported collection of structure-based pharmacophore (68,056 entries) from 8,166 protein−ligand X-ray structures. A diverse set of 157 PDB ligands was profiled using 10 screening protocols on the entire pharmacophore collection, thus generating as many matrices of about 11 million data points. Pharmacophore mapping was compared to another 3-D structure-based method (docking) and to ligand-centric approaches (2-D and 3-D similarity search). In most cases, ligand-based profiling methods outperformed structure-based approaches in their ability to recover the true targets among top-scoring entries. Fine analysis of successes and failures for all methods suggests the design of a hybrid profiling method using the best possible approach as a function of ligand and binding site properties.

METHODS

Generation of Receptor−Ligand Pharmacophores. The Receptor−Ligand Pharmacophore Generation (RLPG) Protocol in Discovery Studio 3.1[38] generates pharmacophore models directly from the receptor−ligand interactions as revealed in the 3-D structures. The RLPG protocol has some notable features: (i) it is fully automated and quickly converts receptor−ligand complexes into pharmacophore models, (ii) it uses adjustable constraints to determine the receptor−ligand interactions, and (iii) it creates all possible pharmacophore combinations, ranks the pharmacophores by decreasing selectivity score, and returns the top-ranked ones. The overall procedure is briefly described as follows.

In the first step, pharmacophore features of the ligand are identified. Six standard pharmacophore features are considered: hydrogen bond acceptor (HBA), hydrogen bond donor (HBD), positive ionizable (PI), negative ionizable (NI), hydrophobic (HYD), and ring aromatic (RA). In the second step, the algorithm prunes all features that do not match the protein−ligand interactions using adjustable topological rules.[39]

Hydrogen Bond Donors/Acceptors. Hydrogen bonds between the protein and ligand are identified, with a default distance of 3.0 Å between heavy atoms. If the enumerated HBA or HBD feature matches the hydrogen bond interaction between the receptor and ligand, it is retained. All others are removed.

Hydrophobic. Hydrophobic features on the ligand are retained if they are within 5.5 Å of the centroid of a hydrophobic residue (Ala, Cys, Ile, Leu, Met, Phe, and Val).

Positive and Negative Ionizable. If an opposite charge center is found on the protein side within 5.6 Å, the feature is retained. All others are removed.

Ring Aromatic. This feature is retained if an aromatic ring is found on the protein and is 2.5 Å away from the projection point of the RA on the ligand.

To construct pharmacophore models that are both sensitive and selective, all combinations of three to six features pharmacophores are enumerated and ranked by decreasing selectivity. Only the top 10 models are selected. There are two options for adding steric constraints to the pharmacophores: shape constraints or excluded volumes. The ligand is used as a template when creating a shape. When adding excluded volumes, an exclusion sphere is added for each neighboring residue. The size of the exclusion sphere is proportional to the number of neighboring protein atoms within a 4−5 Å distance range.

Genetic Function Approximation (GFA) Model for Estimating Pharmacophore Selectivity. The selectively of a pharmacophore model depends on the number of features,

the feature types and their 3-D arrangement. The selectivity is proportional to the number of hits retrieved upon searching a diverse 3-D database. The more hits from the 3-D search the less selective the model. However, it is not practical to screen thousands of pharmacophore candidates using a 3-D database search. Therefore, a mathematical model was created to predict the number of hits rather than performing the search itself. The model was built using default settings of the GFA algorithm[40] embedded in Discovery Studio. Some details for building the GFA model are given as follows.

Druglike Diverse Database. A Catalyst 3-D database of 5,390 druglike diverse ligands was generated in Discovery Studio (default settings of the Build 3D database protocol). The drug-like data set consisted of 3,000 drug-like compounds randomly selected from the BioinfoDB database[41,42] and 2,390 selected from the CAPDiverse database in Discovery Studio.

Diverse Pharmacophore Models. A total of 1,544 pharmacophore models are generated from 200 nonredundant sc-PDB protein−ligand complexes with 2−8 features. Each pharmacophore is used to search the Druglike Diverse database, and the logarithmic value of the number of hits is used for training the GFA model. For pharmacophores with 2−5, 6, 7, and 8 features, the logarithmic value for zero hits is approximated as ln(0.3), ln(0.1), ln(0.03), and ln(0.01), respectively.

Descriptors. Two types of descriptors are used to describe a pharmacophore model: (i) feature set descriptors and (ii) feature−feature distance descriptors. Ten descriptors (number of features, count of certain feature types) were used to specify the feature set of a pharmacophore. The remaining 210 descriptors are related to the feature locations. For each pair of feature types, the feature−feature distance is put into a distance bin (1−10), with a bin size of 2.0 Å. The distance bin count is used as the descriptor value. For instance, descriptors Desc11−Desc20 are used for HBA−HBA distances. Desc11 is the number of HBA−HBA distances in the range of 0−2.0 Å. Desc12 is the number of HBA-HBA distances in the range of 2.0−4.0 Å, and so on. Distances greater than 20.0 Å are counted in the last bin, i.e., Desc20. Similar descriptors are defined for the other types of feature−feature distances. The descriptors and the corresponding feature−feature distance types are shown in Table 1 of the Supporting Information.

GFA Model. Ten GFA models were created using the pharmacophore descriptors to predict for each of the 1,544 diverse pharmacophores the logarithmic value of the number of obtained hits (GFAscore). The best GFA model contains six terms and exhibits an R^2 value of 0.881 (Figure 1 of the Supporting Information). Selectivity is derived from the GFA score using the following equation:

$$\text{Selectivity} = 11 - \text{GFAscore}.$$

The constant of 11 ensures that the selectivity scores will be positive in nearly all cases.

The PharmaDB Pharmacophore Data Set. The *RPLG* algorithm was applied to 8,166 protein−ligand complexes from the sc-PDB database (release 2010).[4] Proteins and ligands were downloaded from the sc-PDB Web site[43] in mol2 file format with formal charges and used as input files by the pharmacophore generation protocol. Three to six features were required for each pharmacophore model, and up to 10 pharmacophores per complex were created. Default settings were used to assign pharmacophoric features on both the ligand and the receptor, as well as for detecting receptor−ligand interactions on the fly. No ligand shape information was explicitly defined, only receptor-based exclusions spheres were included in all models. A total of 7,687 out of the starting 8,166 sc-PDB complexes yielded at least one valid pharmacophore. Altogether, the PharmaDB collection totals 68,056 pharmacophores (chm files) from 2,556 different targets and 3,916 unique ligands.

The sc-PDB Diverse Ligand Set. All sc-PDB binding sites were clustered according to their pairwise similarity computed by the FuzCav method.[44] Briefly, FuzCav converts 3-D atomic coordinates into a vector of 4,834 integers reporting counts of all possible pharmacophoric feature triplets (H-bond acceptor, H-bond donor, positive ionizable, negative ionizable, aromatic, hydrophobic) from binding site-lining residues. The full similarity matrix was converted into a distance matrix, and a hierarchical clustering (average linkage) was then applied in addition to a stopping criterion (Distance >0.84) for the cluster agglomeration. A total of 1,416 binding sites clusters containing 4,228 different ligands were defined. This ligand set was filtered for druglikeness with Filter[45] (see filtering rules in Table 2 of the Supporting Information) to yield a total of 939 unique druglike sc-PDB ligands. A single ligand was randomly chosen for each populated cluster at the condition that the corresponding targets were nonredundant. Finally a total of 182 sc-PDB ligands was obtained and supplemented by 18 ligands from the Astex Diverse Set[46] not present in the sc-PDB. To remove chemical similarity, all 200 remaining ligands were compared using ECFP4 circular fingerprints[47] and kept if dissimilar enough (Tanimoto coefficient <0.7) from all other compounds of the set. This procedure led to a total of 157 unique ligands (sc-PDB Diverse Ligand Set), available in 2-D sd and 3-D mol2 file formats for download from our Web site.[48]

Computational Ligand Profiling. The 157 ligands from the sc-PDB Diverse Set were profiled against the 2,556 targets of the PharmaDB collection (7,687 sc-PDB entries) using four different virtual screening methods.

2-D Similarity Search. The similarity of the query ligand to the starting 3,916 sc-PDB ligands was computed from ECFP4 fingerprints in PipelinePilot[49] and the corresponding targets ranked by decreasing Tanimoto coefficient. The highest similarity value was kept for every sc-PDB target.

3-D Similarity Search. A conformer database (3-D sd file format) was generated using default FAST settings of the *Conformation Generator* component in Pipeline Pilot from each query ligand (2-D sd file) and compared to all sc-PDB ligands (X-ray structure) with ROCS.[50] The corresponding targets were ranked by decreasing Comboscore of their cognate ligands. The highest value was kept for every sc-PDB target.

PharmaDB Pharmacophore Search. The above-described conformer database of query ligands was mapped to all PharmaDB pharmacophores using default settings of the *citest* executable in Discovery Studio. Only the best mapping was kept for each query ligand, and the maximum number of omitted features was set to −1 in both rigid and flexible fitting mode. For every query ligand, targets were ranked by decreasing fitvalue. A second score, the adjusted fitvalue was computed as follows.

$$\text{Adjusted Fitvalue} = (\text{Fitvalue} \times M)/T$$

where M is the number of mapped features, and T is the number of total features in the pharmacophore model. This score will provide a little correction if the conformational sampling of the input ligand was not sufficient to get an optimal mapping to pharmacophore features of the model.

Docking. The 157 query ligands were docked to the 7,687 sc-PDB binding sites using default settings of Surflex[51] (v2.412) and Plants[52] (v.1.2) programs. A maximum of 10 docking poses was saved for every ligand by each program according to their native scoring function (pK_D for surflex, ChemPLP for Plants). All poses were rescored using the FingerPrintLib program,[53] which converts protein−ligand coordinates into molecular interaction fingerprints. Two fingerprints were computed: the first one registers eight interactions (one bit/interaction) per binding site residue (hydrophobic, aromatic, H-bond, ionic, and metal complexion), and the second stores only information from polar interactions (H-bond, ionic, and metal complexion). Similarity of both fingerprints to the X-ray sc-PDB complex was expressed by Tanimoto coefficients (Tc1, Tc2). Only poses with a Tc1 value higher than 0.6 and a Tc2 value higher than 0.2 were kept. Remaining poses were then ranked by decreasing docking score.

RESULTS AND DISCUSSION

PharmaDB Pharmacophore Collection. A collection of 68,056 pharmacophore models has been automatically generated from 7,687 out of the starting 8,166 sc-PDB complexes (conversion rate of 94%). Several reasons of failure could be identified (Table 1). In most of the cases, no pharmacophore model was outputted because of a limited number of pharmacophore features (less than 3). The second major reason was the close proximity of some features that did not respect a minimal interdistance threshold (1.0 Å) and were therefore removed. Last, very few entries contains either valence errors for the bound ligand or did not lead to any feature mapping and should consequently be removed from the next sc-PDB version. A total of 54% of the pharmacophores have the maximum requested number of six features, 18% have five features, 15% have four features, and 13% have three features (Figure 1). Therefore, a very large majority of sc-PDB entries is described by the upper limit of 10 different pharmacophore models. As expected, the selectivity is dependent on the number of features, complex pharmacophores being more selective than simpler ones (Figure 1). The PharmaDB pharmacophore collection describes the interaction of 3,916 unique ligands with 2,556 unique targets and is by far the largest repository of pharmacophores reported to date.

Table 1. Failures in Processing sc-PDB Entries

Failure	Number of cases
No feature mapping	3
Minimal feature−feature distance criteria not fulfilled	79
No pharmacophore maps the ligand	11
Less than 3 features	372
Invalid valence	14

The sc-PDB Ligand Diverse Set. In order to generate the minimum amount of redundancy in the protein−ligand computational matrix, a diverse set of ligands was extracted from the sc-PDB on the basis of the 3-D diversity of the binding sites onto which they bind. The general idea was to select unique ligands binding to dissimilar cavities. Examination of standard molecular properties confirms that selected ligands are druglike, not biased toward unintended property ranges, and chemically dissimilar (Figure 2 A−G). Out of the 157 ligands of the sc-PDB Diverse Set, 130 (83%) have been cocrystallized with a single target in the sc-PDB, 19 have two different targets, six have three different targets, one ligand has four different targets, and one ligand has five different targets (Figure 2H; see full list of targets in Table 3 of the Supporting Information). In total, 165 unique targets are addressed by the Diverse Set with a functional annotation, according to the Enzyme Commission number, similar to that of the full sc-PDB database (Figure 2I).

For further comparing the merits of various profiling methods, the set was divided in two parts: Set 1 includes 29 ligands with targets present in multiple sc-PDB entries and Set 2 describes 128 compounds addressing a target present in a single sc-PDB entry. Because we decided to restrict chemical space to sc-PDB ligands (of known binding mode to their targets), Set 1 ligands can therefore be profiled with either ligand or structure-centric methods, whereas Set 2 ligands can only be profiled by structure-based approaches.

Comparison of Ligand-Based and Structure-Based Profiling Methods (Ligand Set 1). Two ligand-based and two structure-based methods were used to profile the 128 ligands of the Diverse Set 1 against the 7,687 sc-PDB entries (Table 2). For the target-ligand-based pharmacophore approach, two scoring schemes were used according to the ligand-to-pharmacophore fitting procedure (rigid or flexible). For both docking programs (Surflex, Plants), either the native docking score (pK_D for Surflex, ChemPLP for Plants) or a combination of interaction fingerprint similarity and docking score (see Methods) was utilized to rank order sc-PDB targets. Altogether, 10 profiling protocols were then used to yield as many protein−ligand interaction matrices (Table 2). In the current analysis, only protein−ligand complexes registered in the sc-PDB were considered as true positives, although still unknown cross-reaction of some ligands with sc-PDB targets are theoretically possible. The first criteria to estimate the performance of each profiling method was to compute the rank of the true target for every ligand of the Diverse Set. A more qualitative analysis of the target fishing performance was realized by classifying the profiling in three categories: successful (rank of the true target $\leq$25), ambiguous (25 < rank $\leq$50), and failed (rank >50). In case a compound binded to more than one target (27/157 ligands), the highest-ranked target was considered. The threshold of 25 was chosen both on theoretical and practical considerations because it corresponds to the top 1% ranking targets and a target list of manageable size for experimental confirmation.

When all profiling methods can be compared (29 ligands of Set1), ligand-based methods clearly outperform structure-based approaches both quantitatively and qualitatively (Figures 3, 4; Table 4 of the Supporting Information). It should be stated at this point that sc-PDB entries cocrystallized with the ligand to profile were not considered in the analysis below. A 2-D similarity search, the fastest method evaluated in the current study, was slightly better (median rank of the true target = 3, success rate = 76%) than 3-D pharmacophoric shape matching (median rank = 5, success rate = 72%). Although to be expected, this observation is appealing because only a very restricted ligand space (3,916 sc-PDB ligands) was considered, and a simplest nearest-neighbor approach was utilized to rank the corresponding targets. The propensity of structural biologists to cocrystallize multiple ligands of the same chemical series certainly provides a slight bias in our observation. However, this bias would be largely counter-balanced by an extension of ligand space to larger bioactivity databases (e.g., CHEMBL).[5] The two structure-based approaches (pharmaco-

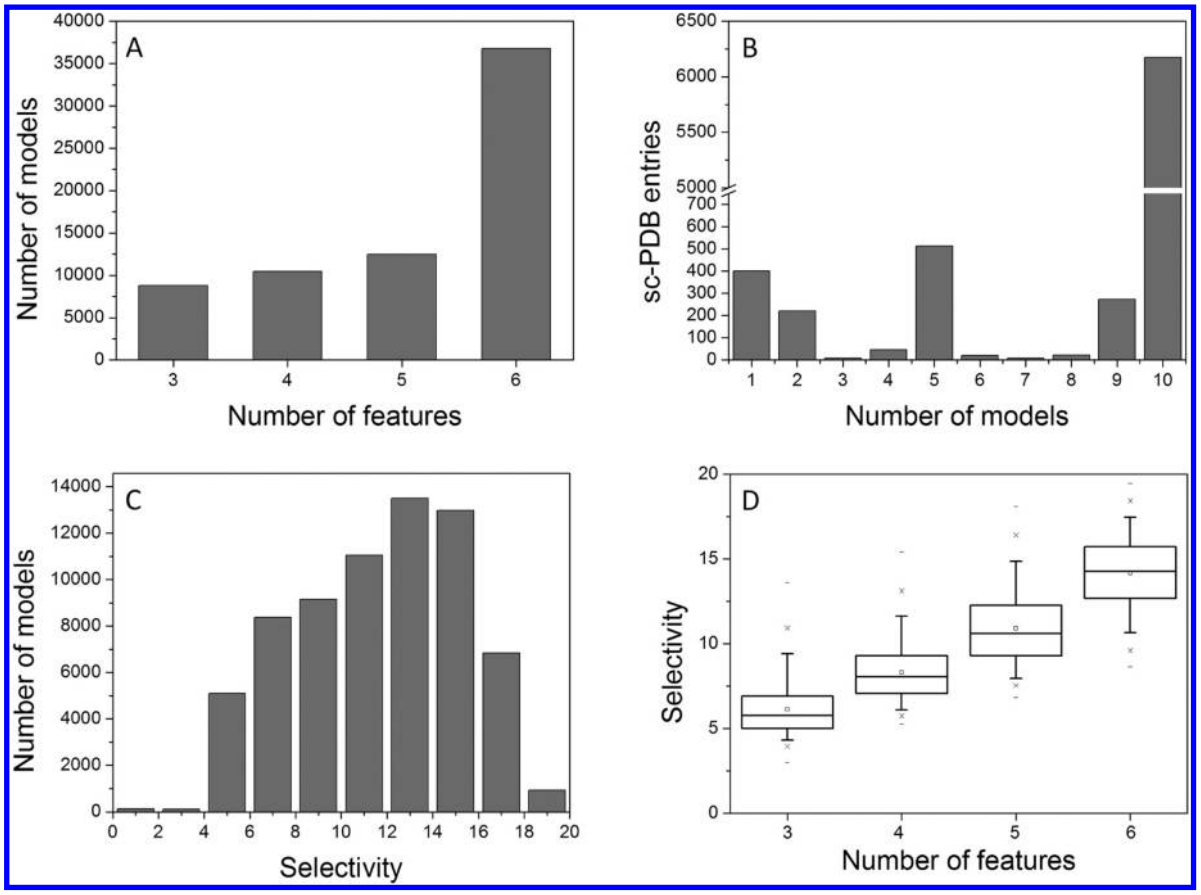

Figure 1. The PharmaDB collection of pharmacophores. (A) Distribution of the number of features by pharmacophore model. (B) Distribution of the number of pharmacophore models by sc-PDB entry. (C) Distribution of the selectivity value of pharmacophore models. (D) Box-and-whisker plot of selectivity value distributions according to the number of features in pharmacophore models. The box delimit the 25th and 75th percentiles; the whiskers delimit the 5th and 95th percentiles. Median and mean values are indicated by a horizontal line and an empty square in the box. Crosses delimit the 1% and 99th percentiles, respectively. Minimum and maximum values are indicated by a dash.

phore search, docking) were significantly less accurate than ligand-centric methods but could be improved by using customized scoring functions (Figure 4). Using an adjusted fitvalue improved the success rate of both rigid (+17%) and flexible (+7%) pharmacophore matches with respect to the native fitvalue. A similar observation was done when docking poses, whatever the program (Surflex, Plants), were first postprocessed by interaction fingerprint similarity to the native X-ray structure and then ranked by decreasing docking score (+21% and +14% increase in success rate for Surflex and Plants, respectively). Flexible pharmacophore match, although much more computer-demanding, was only marginally better (median rank = 12, 66% of success) than rigid fit (median rank = 17, 62% of success). The current study confirms that unbiased docking is the worst performing method (44 and 20% of success for Surflex and Plants, respectively) due to the inaccuracy of standard scoring functions. We believe that this observation is independent of the docking tool as similar evidence were recently reported for the Glide program with obvious interprotein scoring noises (e.g., some targets classes being systematically underestimated, others being systematically overestimated).[54] Fortunately, correcting the biases induced by the native scoring functions by a topological filter (removing poses leading to interaction fingerprints dissimilar to that obtained from known complexes) drastically enhance the performance of both docking tools (+21% and +14% for Surflex and Plants, respectively). Surflex was better suited than Plants for the current profiling set and achieved a performance comparable to that obtained with the best receptor−ligand pharmacophore matching protocols (Figure 4). We could not identify any rule relating profiling accuracy to the binding affinity for the true target. For example, all methods correctly identified S-adenosyl-L-homocysteine hydrolase as a target of 3-deaza adenosine (HET code AD3), although its IC_{50} is reported to be close to 20 μM. Alternatively, no method was able to identify the true target of nanomolar ligands (e.g., I84 aldose reducatase inhibitor, P34 chomera toxin inhibitor; Table 4 of the Supporting Information).

Comparison of Structure-Based Profiling Methods (Ligand Set 2). A deeper comparative analysis of structure-based pharmacophore and docking profiling methods could be drawn from the profiling results of 128 compounds from Set 2, cocrystallized with a single sc-PDB target. Obtained results were in general agreement with the above-reported data for Set 1 compounds, however with some variations (Figure 5): (i) flexible fitting did not ameliorate the accuracy of pharmacophore matching with respect to the simpler rigid fitting procedure, (ii) use of an adjusted fitness value increases the success rate of the pharmacophore-based profiling (+13% for rigid fitting; +8% for flexible fitting), (iii) unbiased docking with native scoring functions is not suited for ligand profiling, and (iv) postprocessing docking poses by interaction fingerprint similarity to the native X-ray structure dramatically enhances

dx.doi.org/10.1021/ci300083r | J. Chem. Inf. Model. 2012, 52, 943−955

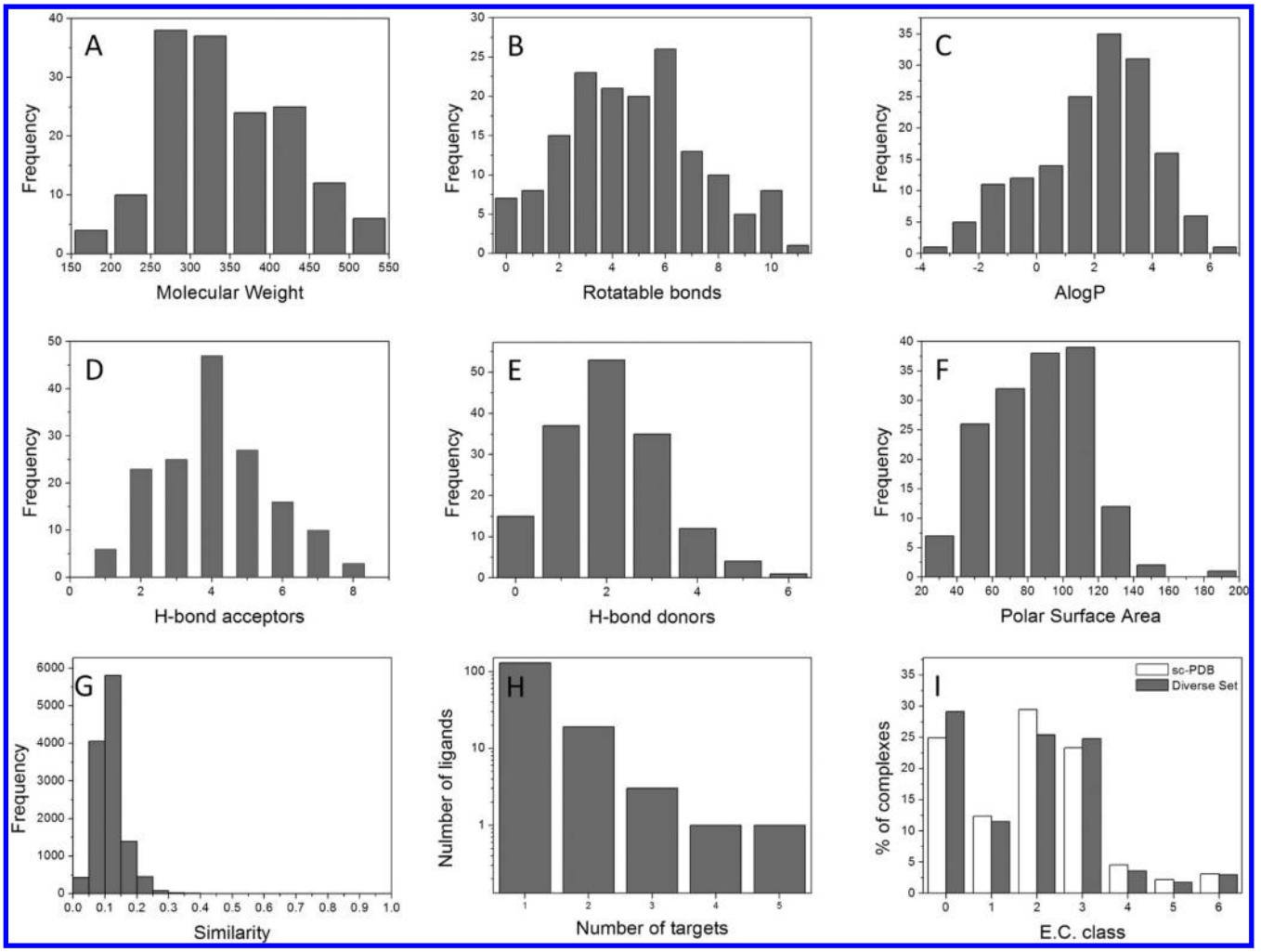

Figure 2. Properties of 157 druglike ligands of the sc-PDB Diverse Set. (A) Molecular weight distribution. (B) Number of rotatable bonds. (C) AlogP, computed logP. (D) Hydrogen-bond acceptor count; (E) Hydrogen-bond donor coun. (F) Polar surface area, Å^2. (G) All-against-all ligand similarity expressed by a Tanimoto coefficient on ECFP4 fingerprints. (H) Distribution of sc-PDB targets among the Diverse Set. (I) Functional annotation of targets by Enzyme Commission (EC) number: 0, no EC number; 1, oxidoreductases; 2,transferases; 3, hydrolases; 4, lyases; 5, isomerases; and 6, ligases.

Table 2. Computational Ligand Profiling Protocols

Protocol	Method	Scoring
Ligand-based		
ECFP4	2-D similarity	Tanimoto coefficient
ROCS	3-D similarity	Comboscore
Structure-based		
Rigid1	Rigid fit to pharmacophore	FitValue
Rigid2	Rigid fit to pharmacophore	Adjusted FitValue
Flex1	Flexible fit to pharmacophore	Fitvalue
Flex2	Flexible fit to pharmacophore	Adjusted Fitvalue
Surflex1	Docking	pk_D
Surflex2	Docking	IFP + pK_D
Plants1	Docking	ChemPLP
Plants2	Docking	IFP +ChemPLP

the success rate of docking-based profiling (+28% for Surflex, +26% for Plants).

For this data set, pharmacophore-based profiling was clearly superior to docking-based protocols. This statement should however be considered with caution because each compound was matched to a pharmacophore derived from its own interactions with the true target (self-matching), which was not the case for Set 1 compounds. The performance of pharmacophore matching is therefore overestimated. Anyway, the very good median ranks indicate that the automatically derived pharmacophore queries are very specific. Surprisingly, rigid fitting was found to be slightly superior to flexible fit (Figure 5B), which might be due to the different ways ligand atoms may overlap excluded volumes in both fitting procedures.

Ligand-Dependent Performance of Profiling Methods. Analysis of the two profiling maps (Tables 4 and 5 of the Supporting Information) shows strong ligand dependencies for all profiling methods. Up to now, we have just analyzed the global performance of several profiling protocols, and current data clearly advise the usage of ligand-based 2-D similarity methods whenever feasible. However, using a single profiling method for all ligands is not an optimal strategy. From here on, we will examine peculiar cases where a single protocol was successful, analyze the reasons for such behaviors, and propose a rationale for prioritizing the best possible method according to the protein−ligand context.

When To Use 2-D Similarity Search. Out of the 29 profilings for which 2-D similarity search could be applied, only

dx.doi.org/10.1021/ci300083r | J. Chem. Inf. Model. 2012, 52, 943−955

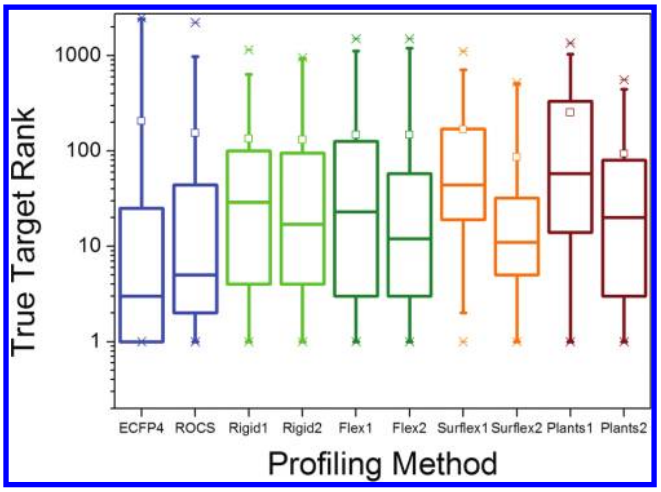

Figure 3. Comparative evaluation of 2 ligand-based and 8 structure-based protocols (see description in Table 2) for profiling 29 ligands (Set 1) against 2,556 unique sc-PDB targets. Box-and-whisker plot of the distribution of true target ranks. The box delimit the 25th and 75th percentiles; the whiskers delimit the 5th and 95th percentiles. Median and mean values are indicated by a horizontal line and an empty square in the box. Crosses delimit the 1% and 99th percentiles, respectively. Minimum and maximum values are indicated by a dash.

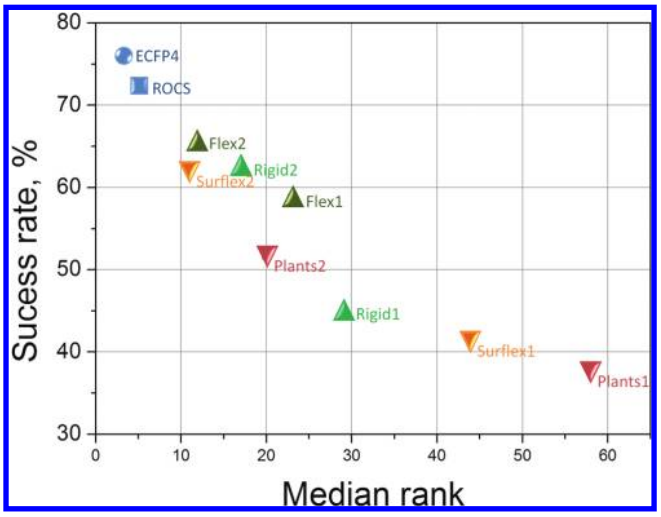

Figure 4. Profiling success rate of 29 ligands (Set 1) as a function of the true target median rank for 10 computational profiling protocols.

five of them failed to recover the true target among the first 50 scoring entries (Table 4 of the Supporting Information). Four of these failures (ET, P34, PVB, RNP) are simply due to an insufficient number of reference ligands (<10) for the cognate targets. The last ligand (I84) that could not be correctly profiled is a human aldose reductase (AR) inhibitor. Although 25 other AR inhibitors were present in the sc-PDB, none of them was similar enough to the query. AR is an enzyme whose 3-D structure is particularly flexible and offers multiple binding modes to chemically diverse inhibitors.[55] Such a behavior is however expected to be more the exception than the rule.

In their Similarity Ensemble Approach (SEA), Keiser et al. only considers targets with more than 100 different ligands.[17,18] Although our data set is very limited, we think that a lower threshold is enough to guarantee an accurate profiling. Thirteen out of the 14 proteins of Set 1 ligands, having more than 20 cocrystallized ligands, were indeed recovered among the top 1%

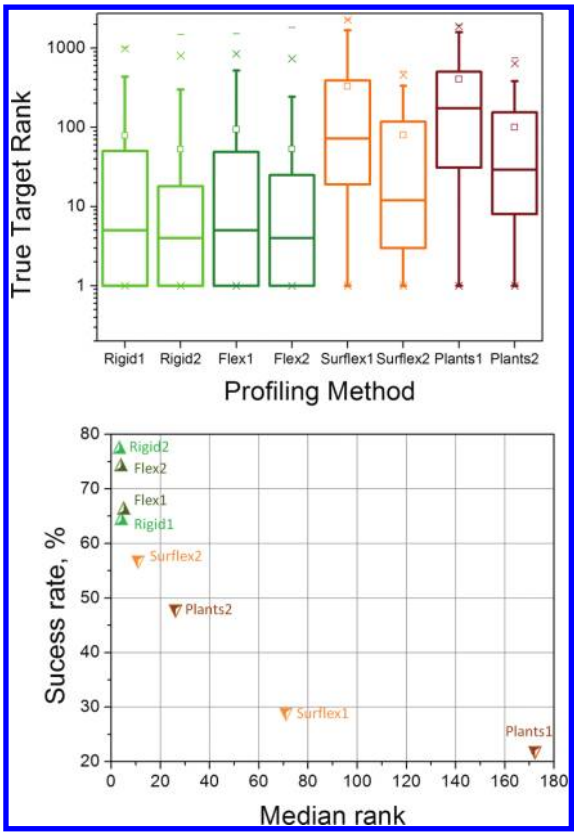

Figure 5. Comparative evaluation of 2 ligand-based and 8 structure-based protocols (see description in Table 2) for profiling 128 ligands (Set 2) of against 2,556 unique sc-PDB targets. (A) Box-and-whisker plot of the distribution of true target ranks. The box delimit the 25th and 75th percentiles; the whiskers delimit the 5th and 95th percentiles. Median and mean values are indicated by a horizontal line and an empty square in the box. Crosses delimit the 1% and 99th percentiles, respectively. Minimum and maximum values are indicated by a dash. (B) Profiling success rate ligands as a function of the true target median rank for 10 computational profiling methods.

scoring entries. We therefore recommend the usage of ligand-based 2-D similarity methods whenever possible, which means when enough different ligands (>20) can be used as references to annotate a target.

In addition to this general consideration, we describe here peculiar ligand and binding site properties favoring exclusively one of the four virtual profiling methods used in the current study.

For example, profiling of the pantothenate kinase inhibitor PAU is only successful with a 2-D ligand similarity search (Table 4 of the Supporting Information). PAU is an acyclic small molecular-weight ligand (MW = 218) highly buried (96%) upon binding to its target (PDB entry 3af0, Figure 6A,B). Its closest sc-PDB ligand with respect to 2-D ECFP4 fingerprints is its phosphate analogue PAZ, another pantothenate kinase inhibitor (PDB entry 3aez; Tc = 0.63; Figure 6C). The 2-D similarity is however not concomitant with 3-D similarity as estimated by ROCS overlay of flexible PAU to rigid PAZ because of the additional presence of an additional polar phosphate group in PAZ (Comboscore = 0.948; Figure 6D). The corresponding true target is thus badly ranked (rank 293) in the ROCS-derived target list. The closest ligand to PAU in our 3-D descriptor space is another low molecular-weight inhibitor (BTW, MW = 209) of a totally different target

dx.doi.org/10.1021/ci300083r | J. Chem. Inf. Model. 2012, 52, 943−955

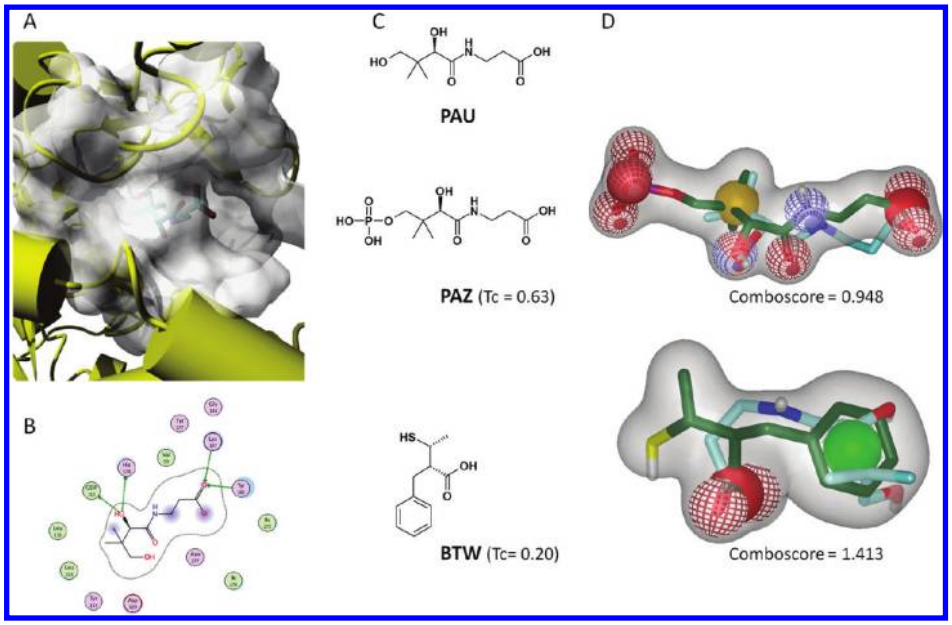

Figure 6. Profiling pantothenic acid (HET code = PAU). (A) X-ray structure of PAU (sticks) in complex with pantothenate kinase (white surface, pdb entry 3af0). (B) Schematic 2-D diagram of protein−ligand interactions. Apolar and polar binding site residues are circled in green and violet, respectively. Hydrogen-bonds are indicated by green arrows. Accessible ligand atoms are enclosed by a cyan dot. (C) Chemical structures of PAU, PAZ, and BTW ligands. 2-D similarity to PAU is indicated in brackets. (D) ROCS overlay of PAU (cyan carbon atoms) with PAZ and BTW (green carbon atoms). Oxygen, nitrogen, and sulfur atoms are colored in red, blue, and yellow, respectively. The shape of templates (PAZ, BTW) is displayed by a white surface. Pharmacophoric features are displayed by balls (H-bond acceptor, red mesh dot; H-bond donor, blue meshed dot; negative ionizable, red solid dot, hydrophobe, yellow dot; and ring, green dot).

(carboxypeptidase A1, PDB entry 3i1u) that better matches in 3-D space (Comboscore = 1.413) although the pairwise 2-D similarity is low (Tc = 0.19; Figures 6C,D). Because of its low molecular weight and high polarity, PAU is found to fit many structure-based pharmacophores and to dock many binding sites, therefore explaining the failure of all structure-based protocols to correctly profile this compound. We therefore suggest profiling, whenever feasible, polar fragment-like compounds with 2-D similarity search methods.

When To Use 3-D Similarity Search. In a single profiling example (ET, Table 4 of the Supporting Information), 3-D similarity search was the only method capable of recovering the true targets among the top 1% scoring entries. Ethidium (ET) is a polyaromatic compound (AlogP = 4.20) binding to the transcriptional regulatory protein Qacr mainly through deeply buried apolar and aromatic pi−pi interactions. A single hydrogen-bond to the binding site is observed (Figure 7A, B). The 2-D similarity of ET to the six other Qacr sc-PDB ligands is low (Tc in the 0.09−0.14 range, Figure 7C), although 3-D ROCS similarity to one of these ligands (MGR) is high (Figure 7D). The shape and three ring aromatic features of MGR are well matched by ligand ET. This ligand matches many other pharmacophores dominated by apolar and aromatic features and exhibiting low specificity values. Docking is also not suited for profiling such compounds where interactions are not directional and therefore badly scored. We thus recommend 3-D similarity search for profiling hydrophobic compounds exhibiting few hydrogen-bond donors/acceptors.

Receptor−Ligand Pharmacophore versus Docking-Based Profiling. Ligand mapping to a receptor−ligand-based pharmacophore can be considered as a variant of molecular docking in which the estimation of protein−ligand interactions is not quantified by a binding energy but a fitness to a topological description. In order to delineate differences in both approaches, we identified ligands profiling cases where rigid pharmacophore search (Rigid2 protocol) was successful but both docking-based protocols (Surflex2, Plants2) failed, and vice versa. Scores used in this comparison were the adjusted fitvalue for the rigid pharmacophore method and the docking score preprocessed by interaction fingerprints for docking as mentioned above.

For nine ligands, only pharmacophore search was successful, whereas eight cases could be reported in which only docking-based approaches were efficient in recovering the true target among the top 1% scoring targets. Examining the molecular properties of both ligand sets show some clear tendencies: ligands suited for docking-based profiling are more polar (higher number of hydrogen-bond donors, lower clogP, and higher polar surface area) than ligands suited for receptor−ligand-based profiling (Figure 8). Interestingly, the properties of the corresponding binding sites matched the above-reported ligands properties. Targets recovered by docking exhibit binding cavities that are more buried, polar, and smaller than those recovered by pharmacophore search (Figure 9). This observation corroborated most benchmarks, indicating that this method is highly sensitive to the directionality of protein−ligand interactions (e.g., hydrogen-bond count), cavity size, and polarity.[56] Receptor−ligand-based pharmacophore search, which can be considered as a constrained docking in which positional and pharmacophoric constraints have been automatically derived, does not suffer from this drawback.

We next compared the quality of the poses generated by either pharmacophore match or docking using the previously identified best profiling procotols for each method (rigid2, flex2, surflex2, plants2; Figure 5). Instead of reporting root-mean square deviations (rmsd) to the X-ray pose, we plotted the similarity of predicted to experimentally observed protein−ligand interactions (Figure 10). The later measure was

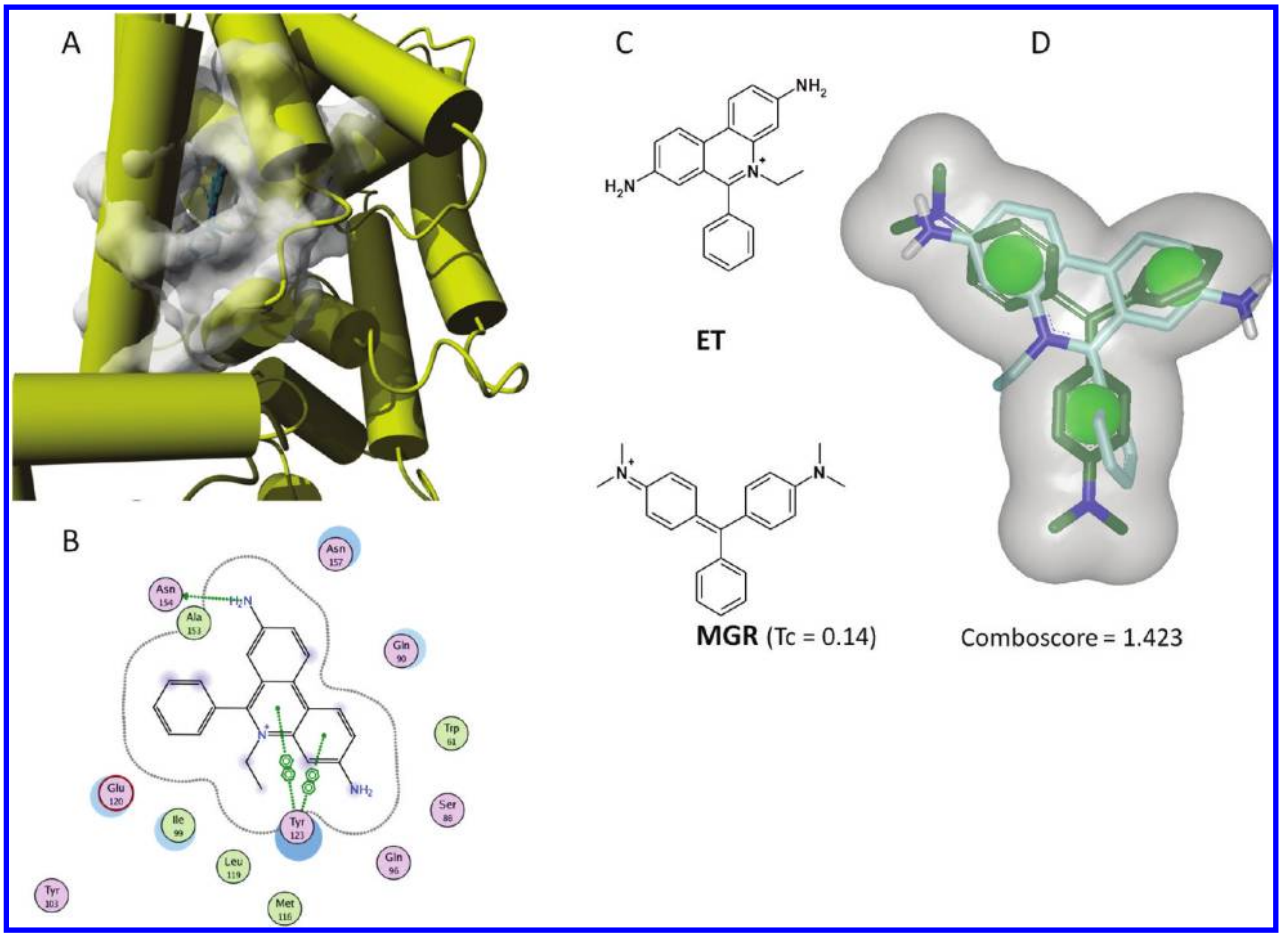

Figure 7. Profiling Ethidium (HET code = ET). (A) X-ray structure of ET (sticks) in complex with the transcriptional regulator protein Qacr (white surface, pdb entry 3br3). (B) Schematic 2-D diagram of protein−ligand interactions. Apolar and polar binding site residues are circled in green and violet, respectively. Hydrogen-bonds are indicated by green arrows. Aromatic pi−pi interactions are displayed by dotted lines. Accessible ligand atoms are enclosed by a cyan dot. (C) Chemical structures of ET and MGR ligands. 2-D similarity of the closest ligand (MGR) to ET is indicated in brackets. (D) ROCS overlay of ET (cyan carbon atoms) with MGR (green carbon atoms). Oxygen and nitrogen atoms are colored in red and blue, respectively. The shape of MGR template is displayed by a white surface. Pharmacophoric features are displayed by balls (ring, green dot).

previously reported to be a much better indicator of pose quality than rmsd.[53] Analysis of the best pose for the 128 Set 2 ligands from pharmacophore match or docking indicates that flexible pharmacophore fitting produces the best poses but with only a marginal superiority to rigid pharmacophore fit and docking (Figure 10). In 80% of the cases, all methods provides an orientation in which about 60% of protein−ligand interactions are conserved (T*c* > 0.6), a threshold considered as acceptable for estimating the quality of a pose.[53] The poorer profiling performance of docking with respect to pharmacophore search (Figure 5) is therefore only attributable to scoring and not to insufficient conformational sampling of the ligand. Interestingly, flexible fit to a receptor−ligand pharmacophore does not provide a substantial advantage to the rigid fitting procedure, although the later is much faster (Table 3, Figure 2 of the Supporting Information). In conclusion, receptor−ligand pharmacophore search can be considered as a reliable and fast alternative to molecular docking.

We recommend this methodology for profiling targets for which few ligands but a 3-D structure is available, with the exception of profiling polar ligands to small, polar, and buried active sites for which molecular docking is preferable.

■ CONCLUSIONS

In this work, we describe a fully automated method to generate 3-D pharmacophore queries from protein−ligand X-ray structures, with an estimation of pharmacophore selectivity based on the number anticipated druglike hits. The protocol was applied to the sc-PDB data set of protein−ligand complexes to generate a database of 68,056 pharmacophores (PharmaDB) describing 2,556 unique targets. This study offered us the opportunity to compare, for the first time, ligand-based and structure-based methods to profile a set of 157 diverse ligands against our panel of targets. When applicable (more than 20 ligands known for each target), ligand-based methods (2-D and 3-D similarity search) clearly outperformed structure-based (pharmacophore, docking) profiling protocols.

Pose accuracy is relatively similar for pharmacophore fitting and docking, whatever the protocol used. Because that accuracy is satisfactory for about 80% of ligands (rmsd < 2 Å, Tc-IFP > 0.6), we conclude that the lower performance of structure-based methods is due to the difficulty of scoring functions to discriminate correct poses from near-native decoys. Failure of scoring functions in docking may be significantly rescued by reranking poses by interaction fingerprint similarity to a known reference (same protein cocrystallized with another ligand),

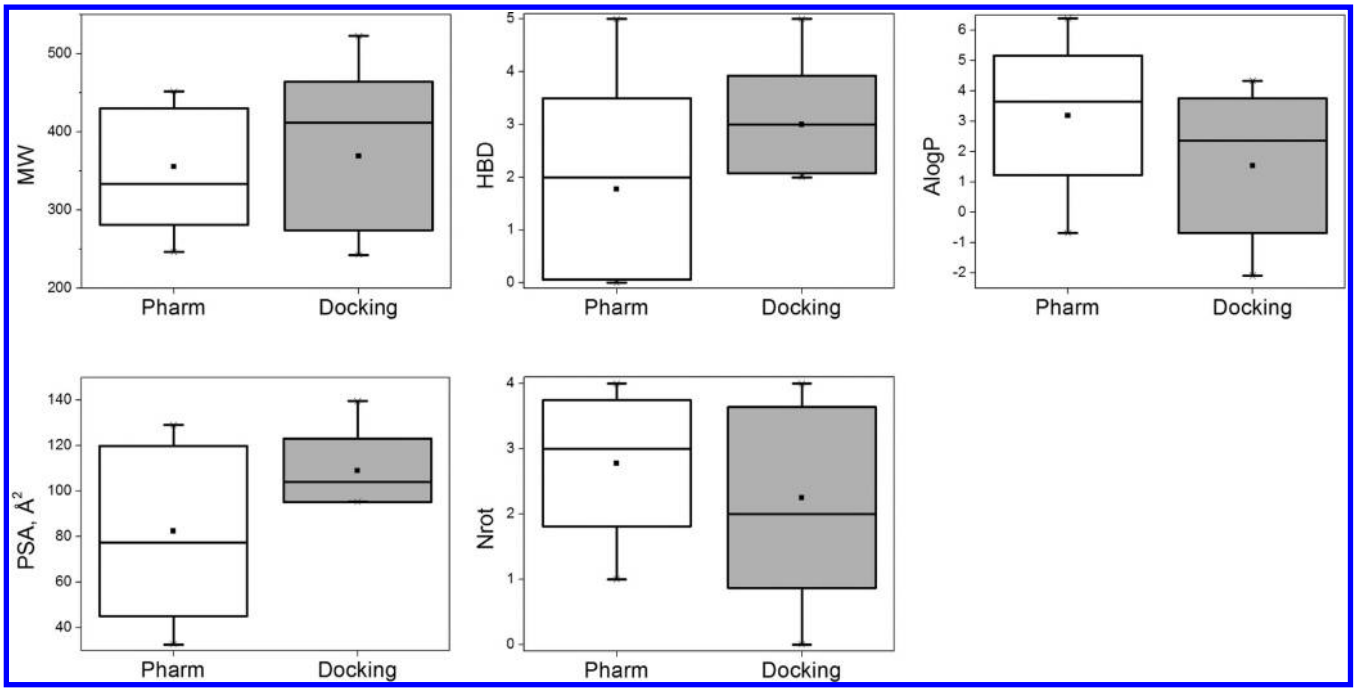

Figure 8. Molecular properties (molecular weight, MW; hydrogen-bond donor count, HBD; predicted log P, AlogP; polar surface area, PSA; and number of rotatable bonds, Nrot), computed by Pipeline Pilot,[49] of ligands whose targets are only recovered by either pharmacophore-based search (Pharm) or docking-based profiling (Docking). The box delimit the 25th and 75th percentiles; the whiskers delimit the 5th and 95th percentiles. Median and mean values are indicated by a horizontal line and an empty square in the box. Crosses delimit the 1% and 99th percentiles, respectively. Minimum and maximum values are indicated by a dash.

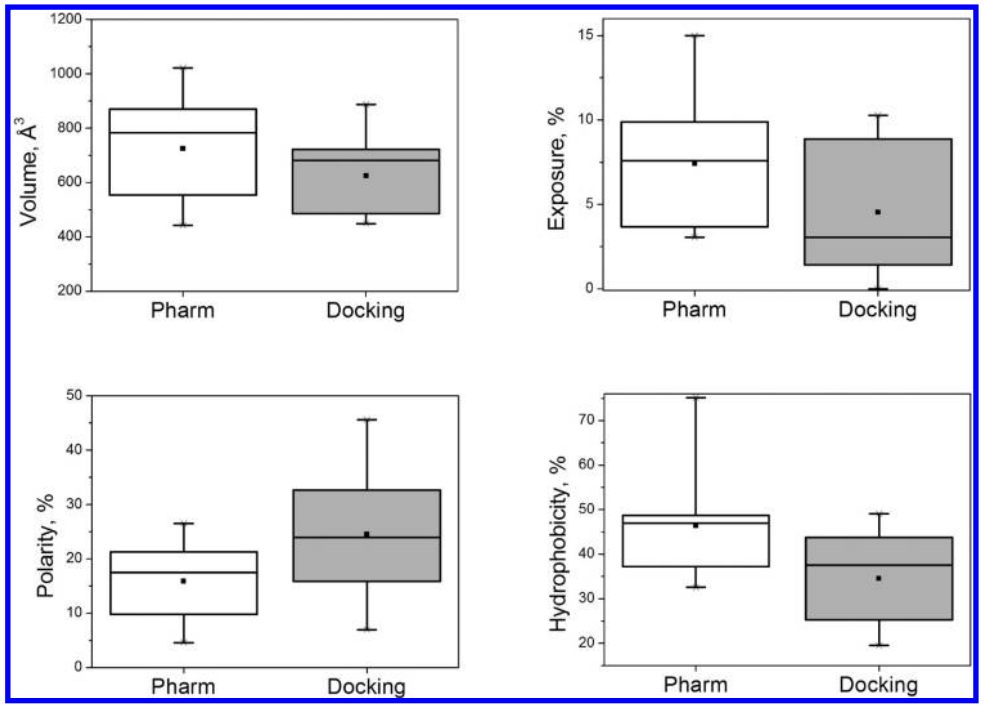

Figure 9. Molecular properties (volume, exposure, polarity, hydrophobicity), computed by VolSite (unpublished in-house program) of binding sites whose targets are only recovered by either pharmacophore-based search (Pharm) or docking-based profiling (Docking). The box delimit the 25th and 75th percentiles; the whiskers delimit the 5th and 95th percentiles. Median and mean values are indicated by a horizontal line and an empty square in the box. Crosses delimit the 1% and 99th percentiles, respectively. Minimum and maximum values are indicated by a dash.

which means that sampling and generating correct poses for most druglike compounds is not the major problem here. Target flexibility is also partially addressed in our application because multiple copies of the same protein bound to various ligands are stored in the sc-PDB and therefore explicitly taken into account in both docking and pharmacophore-based profiling. Multiple reasons remain for explaining inaccurate scoring in docking, all of which have been extensively surveyed in recent reports.[57−59] Some are due to the necessary high throughput requested in preparing a large collection of

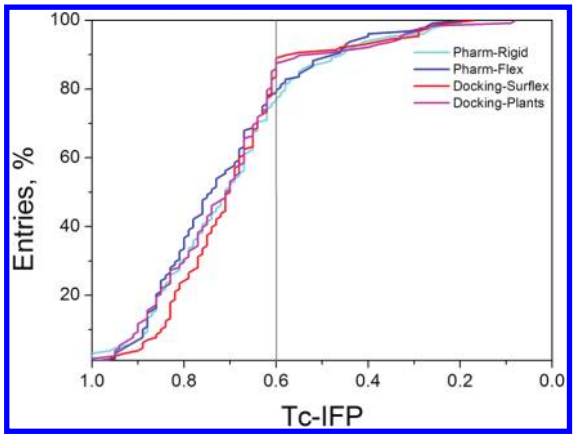

Figure 10. Quality of poses generated by pharmacophore match (rigid, flexible) and docking (Surflex, Plants), for the set of 159 sc-PDB diverse ligands. Similarity to the native X-ray pose is estimated by a Tanimoto coefficient on protein−ligand fingerprints (Tc-IFP).[53] A single pose is retrieved for each method according to its score (adjusted fitvalue for pharmacophore search, pk_D for Surflex, ChemPLP for Plants). Please note that raw docking poses are first processed to discard solutions for which Tc1-IFP is less than 0.6 and Tc2-IFP is less than 0.2 (see Methods). A solid vertical line at Tc1 = 0.6 indicates the threshold for acceptable solutions.

Table 3. Median CPU Time (128 ligands of Set 2) for Profiling One Target by Pharmacophore and Docking Protocols

Method	Method	CPU time, s[a]
Pharmacophore search	Rigid	4
	Flexible	70
Docking	Surflex	25
	Plants	20

[a]CPU time for a 3.16 Ghz Intel Core Duo E 8500 processor with 4 Go RAM.

heterogeneous binding sites for docking: possible inaccurate tautomeric state for some histidines, possible ligand-dependent flip of terminal amide bonds, and omission of ligand-dependent protein-bound water molecules. Some are intrinsic to any docking-based virtual screen: improper handling of dehydration and more globally entropic effects in absence of a slower but more rigorous energy function to rerank binding poses, neglecting the quality of the host protein structure in docking scores, and absence of accurate terms for weaker but sometimes important intermolecular interactions (e.g., weak hydrogen and halogen bonds, quadrupole−quadrupole interactions). Improper treatment of entropic effects and of peculiar protein−ligand interactions also applies to pharmacophore matches. Possible reasons for inaccurate fitting of true actives to structure-based pharmacophores have also been reviewed by many authors.[32,60,61] Of particular concern in our application is the choice and placement of pharmacophoric features (notably hydrophobic features) and of exclusion spheres that are known to strongly influence pharmacophoric matches.

We should point again that the herein derived conclusions have been drawn in a ligand profiling context, which is a particular application of virtual screening. The present conclusion that ligand-centric approaches outperform structure-based methods in profiling is pragmatic and based on existing data at the PDB scale. We clearly demonstrate that the profiling accuracy of all methods is target and binding site dependent. Examining cases of successes and failures suggest the use of hybrid profiling workflows in which all methods should be applied depending on the protein−ligand context. This strategy presents the advantage to be data-driven and to significantly extend the applicability domain of ligand profiling to a wide array of different targets. Additional independent benchmarking studies as well as the increasing availability of high-quality bioactivity data will certainly help in refining the herein derived first conclusions on computational ligand profiling.

■ ASSOCIATED CONTENT

Supporting Information

A plot of GFAscore versus ln(hits), obtained by screening 200 sc-PDB ligands on 1,544 pharmacophore models, a plot of the distribution of profiling time (in seconds) for pharmacophore and docking-based methods, the list of pharmacophore descriptors used for training the GFA model to estimate selectivity, the OpenEye Filter file used to select druglike ligands, the list of HET codes for the 157 ligands of the sc-PDB Diverse Set and their respective targets, and the profiling accuracy of 29 ligands (Set 1) and 128 ligands (Set 2) by 10 computational protocols.This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: rognan@unistra.fr.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank the Conseil regional d'Alsace for funding J.M. The IPHC grid (Strasbourg), CC-IN2P3 (Villeurbanne), and GENCI (Project x2011075024) are acknowledged for providing computational resources. We thank Dr. Esther Kellenberger (UMR 7200, Illkirch) for help with the selection of diverse ligands, and Jérôme Pansanel (IPHC), Pascal Calvat, and David Bouvet (CC-IN2P3) for their friendly and excellent support.

■ REFERENCES

(1) Ekins, S.; Williams, A. J. When pharmaceutical companies publish large datasets: An abundance of riches or fool's gold? *Drug Discovery Today* **2010**, *15*, 812−5.

(2) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B. A.; Bolton, E.; Gindulyte, A.; Bryant, S. H. PubChem's BioAssay Database. *Nucleic Acids Res.* **2011**.

(3) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−42.

(4) Meslamani, J.; Rognan, D.; Kellenberger, E. sc-PDB: A database for identifying variations and multiplicity of "druggable" binding sites in proteins. *Bioinformatics* **2011**, *27*, 1324−6.

(5) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2011**, *40*, D1100−1107.

(6) Rognan, D. Chemogenomic approaches to rational drug design. *Br. J. Pharmacol.* **2007**, *152*, 38−52.

(7) Schneider, G. Virtual screening: An endless staircase? *Nat. Rev. Drug Discov.* **2010**, *9*, 273−6.

(8) Morphy, R. Selectively nonselective kinase inhibition: Striking the right balance. *J. Med. Chem.* **2010**, *53*, 1413−37.

(9) Hopkins, A. L.; Mason, J. S.; Overington, J. P. Can we rationally design promiscuous drugs? *Curr. Opin. Struct. Biol.* **2006**, *16*, 127−36.
(10) Ekins, S.; Williams, A. J.; Krasowski, M. D.; Freundlich, J. S. In silico repositioning of approved drugs for rare and neglected diseases. *Drug Discovery Today* **2011**, *16*, 298−310.
(11) Rognan, D. Structure-based approaches to target fishing and ligand profiling. *Mol. Inf.* **2010**, *29*, 176−187.
(12) Muller, P.; Lena, G.; Boilard, E.; Bezzine, S.; Lambeau, G.; Guichard, G.; Rognan, D. In silico-guided target identification of a scaffold-focused library: 1,3,5-triazepan-2,6-diones as novel phospholipase A2 inhibitors. *J. Med. Chem.* **2006**, *49*, 6768−78.
(13) Yang, L.; Wang, K.; Chen, J.; Jegga, A. G.; Luo, H.; Shi, L.; Wan, C.; Guo, X.; Qin, S.; He, G.; Feng, G.; He, L. Exploring off-targets and off-systems for adverse drug reactions via chemical-protein interactome−clozapine-induced agranulocytosis as a case study. *PLoS Comput. Biol.* **2011**, *7*, e1002016.
(14) Keiser, M. J.; Setola, V.; Irwin, J. J.; Laggner, C.; Abbas, A. I.; Hufeisen, S. J.; Jensen, N. H.; Kuijer, M. B.; Matos, R. C.; Tran, T. B.; Whaley, R.; Glennon, R. A.; Hert, J.; Thomas, K. L.; Edwards, D. D.; Shoichet, B. K.; Roth, B. L. Predicting new molecular targets for known drugs. *Nature* **2009**, *462*, 175−81.
(15) Vidal, D.; Garcia-Serna, R.; Mestres, J. Ligand-based approaches to in silico pharmacology. *Methods Mol. Biol.* **2010**, *672*, 489−502.
(16) Yera, E. R.; Cleves, A. E.; Jain, A. N. Chemical structural novelty: On-targets and off-targets. *J. Med. Chem.* **2011**, *54*, 6771−85.
(17) Keiser, M. J.; Hert, J. Off-target networks derived from ligand set similarity. *Methods Mol. Biol.* **2009**, *575*, 195−205.
(18) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25*, 197−206.
(19) Surgand, J. S.; Rodrigo, J.; Kellenberger, E.; Rognan, D. A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors. *Proteins* **2006**, *62*, 509−38.
(20) Xie, L.; Bourne, P. E. Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 5441−6.
(21) Martin, R. E.; Green, L. G.; Guba, W.; Kratochwil, N.; Christ, A. Discovery of the first nonpeptidic, small-molecule, highly selective somatostatin receptor subtype 5 antagonists: A chemogenomics approach. *J. Med. Chem.* **2007**, *50*, 6291−4.
(22) Defranchi, E.; Schalon, C.; Messa, M.; Onofri, F.; Benfenati, F.; Rognan, D. Binding of protein kinase inhibitors to synapsin I inferred from pair-wise binding site similarity measurements. *PLoS One* **2010**, *5*, e12214.
(23) Kinnings, S. L.; Liu, N.; Buchmeier, N.; Tonge, P. J.; Xie, L.; Bourne, P. E. Drug discovery using chemical systems biology: Repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS Comput. Biol.* **2009**, *5*, e1000423.
(24) van Westen, G. J. P.; Wegner, J. K.; Ijzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *MedChemComm* **2011**, *2*, 16−30.
(25) Wolber, G.; Langer, T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* **2005**, *45*, 160−9.
(26) Yang, L.; Chen, J.; He, L. Harvesting candidate genes responsible for serious adverse drug reactions from a chemical-protein interactome. *PLoS Comput. Biol.* **2009**, *5*, e1000441.
(27) Li, Y. Y.; An, J.; Jones, S. J. A computational approach to finding novel targets for existing drugs. *PLoS Comput. Biol.* **2011**, *7*, e1002139.
(28) Durrant, J. D.; Amaro, R. E.; Xie, L.; Urbaniak, M. D.; Ferguson, M. A.; Haapalainen, A.; Chen, Z.; Di Guilmi, A. M.; Wunder, F.; Bourne, P. E.; McCammon, J. A. A multidimensional strategy to detect polypharmacological targets in the absence of structural and sequence homology. *PLoS Comput. Biol.* **2010**, *6*, e1000648.
(29) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L., 3rd Assessing scoring functions for protein−ligand interactions. *J. Med. Chem.* **2004**, *47*, 3032−47.
(30) Enyedy, I.; Egan, W. Can we use docking and scoring for hit-to-lead optimization? *J. Comput.-Aided Mol. Design* **2008**, *22*, 161−168.
(31) Wang, F.; Liu, D.; Wang, H.; Luo, C.; Zheng, M.; Liu, H.; Zhu, W.; Luo, X.; Zhang, J.; Jiang, H. Computational screening for active compounds targeting protein sequences: Methodology and experimental validation. *J. Chem. Inf. Model.* **2011**, *51*, 2821−8.
(32) Leach, A. R.; Gillet, V. J.; Lewis, R. A.; Taylor, R. Three-dimensional pharmacophore methods in drug discovery. *J. Med. Chem.* **2010**, *53*, 539−58.
(33) Steindl, T. M.; Schuster, D.; Laggner, C.; Langer, T. Parallel screening: A novel concept in pharmacophore modeling and virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 2146−57.
(34) Steindl, T. M.; Schuster, D.; Laggner, C.; Chuang, K.; Hoffmann, R. D.; Langer, T. Parallel screening and activity profiling with HIV protease inhibitor pharmacophore models. *J. Chem. Inf. Model.* **2007**, *47*, 563−71.
(35) Markt, P.; Schuster, D.; Kirchmair, J.; Laggner, C.; Langer, T. Pharmacophore modeling and parallel screening for PPAR ligands. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 575−90.
(36) Liu, X.; Ouyang, S.; Yu, B.; Liu, Y.; Huang, K.; Gong, J.; Zheng, S.; Li, Z.; Li, H.; Jiang, H. PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach. *Nucleic Acids Res.* **2010**, *38*, W609−14.
(37) Rollinger, J. M.; Schuster, D.; Danzl, B.; Schwaiger, S.; Markt, P.; Schmidtke, M.; Gertsch, J.; Raduner, S.; Wolber, G.; Langer, T.; Stuppner, H. In silico target fishing for rationalized ligand discovery exemplified on constituents of Ruta graveolens. *Planta Med.* **2009**, *75*, 195−204.
(38) Discovery Studio v.3.1.0; Accelrys Software, Inc.: San Diego, CA.
(39) Sutter, J.; Li, J.; Maynard, A. J.; Goupil, A.; Luu, T.; Nadassy, K. New features that improve the pharmacophore tools from Accelrys. *Curr. Comput.-Aided Drug Des.* **2011**, *7*, 173−80.
(40) Rogers, D.; Hopfinger, A. J. Application of genetic function approximation to quantitative structure−activity relationships and quantitative structure−property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.
(41) 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*, 1294−303.
(42) Bioinf-DB 11.2, version 2011.1. http://bioinfo-pharma.u-strasbg.fr/bioinfo (accessed June 2011).
(43) sc-PDB. http://bioinfo-pharma.u-strasbg.fr/scPDB (accessed June 2011).
(44) Weill, N.; Rognan, D. Alignment-free ultra-high-throughput comparison of druggable protein−ligand binding sites. *J. Chem. Inf. Model.* **2010**, *50*, 123−35.
(45) Filter v.2.0.2; OpenEye Scientific Software: Santa Fe, NM.
(46) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726−41.
(47) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−54.
(48) sc-PDB Diverse Ligand Set. http://bioinfo-pharma.u-strasbg.fr/labwebsite/downloads/scPDB_DiverseSet.zip.
(49) Pipeline Pilot v.8.5.0; Accelrys Software, Inc.: San Diego, CA.
(50) ROCS v.3.1.2; OpenEye Scientific Software: Santa Fe, NM.
(51) Jain, A. N. Surflex-Dock 2.1: Robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 281−306.
(52) Korb, O.; Stutzle, T.; Exner, T. E. Empirical scoring functions for advanced protein−ligand docking with PLANTS. *J. Chem. Inf. Model.* **2009**, *49*, 84−96.
(53) Marcou, G.; Rognan, D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.* **2007**, *47*, 195−207.

(54) Wang, W.; Zhou, X.; He, W.; Fan, Y.; Chen, Y.; Chen, X. The interprotein scoring noises in glide docking scores. *Proteins* **2011**, *80*, 169−83.
(55) Steuber, H.; Zentgraf, M.; La Motta, C.; Sartini, S.; Heine, A.; Klebe, G. Evidence for a novel binding site conformer of aldose reductase in ligand-bound state. *J. Mol. Biol.* **2007**, *369*, 186−97.
(56) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **2004**, *57*, 225−42.
(57) Schneider, N.; Hindle, S.; Lange, G.; Klein, R.; Albrecht, J.; Briem, H.; Beyer, K.; Claussen, H.; Gastreich, M.; Lemmen, C.; Rarey, M. Substantial improvements in large-scale redocking and screening using the novel HYDE scoring function. *J Comput.-Aided Mol. Des.* **2011**.
(58) Novikov, F. N.; Zeifman, A. A.; Stroganov, O. V.; Stroylov, V. S.; Kulkov, V.; Chilov, G. G. CSAR scoring challenge reveals the need for new concepts in estimating protein−ligand binding affinity. *J. Chem. Inf. Model.* **2011**, *51*, 2090−6.
(59) Smith, R. D.; Dunbar, J. B., Jr.; Ung, P. M.; Esposito, E. X.; Yang, C. Y.; Wang, S.; Carlson, H. A. CSAR benchmark exercise of 2010: Combined evaluation across all submitted scoring functions. *J. Chem. Inf. Model.* **2011**, *51*, 2115−31.
(60) Wolber, G.; Seidel, T.; Bendix, F.; Langer, T. Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discovery Today* **2008**, *13*, 23−9.
(61) Spitzer, G. M.; Heiss, M.; Mangold, M.; Markt, P.; Kirchmair, J.; Wolber, G.; Liedl, K. R. One concept, three implementations of 3D pharmacophore-based virtual screening: Distinct coverage of chemical search space. *J. Chem. Inf. Model.* **2010**, *50*, 1241−7.