Hot-Spots-Guided Receptor-Based Pharmacophores (HS-Pharm): A Knowledge-Based Approach to Identify Ligand-Anchoring Atoms in Protein Cavities and Prioritize Structure-Based Pharmacophores

Caterina Barillari,[†] Gilles Marcou,[‡] and Didier Rognan[*,†]

Bioinformatics of the Drug, UMR 7175 CNRS-ULP (Université Louis Pasteur-Strasbourg I), 74 route du Rhin, B.P. 24, F-67400 Illkirch, France, and Laboratoire d'Infochimie, UMR 7551 CNRS, Université Louis Pasteur-Strasbourg I, 4 rue B. Pascal, F-67000 Strasbourg, France

Received February 21, 2008

The design of biologically active compounds from ligand-free protein structures using a structure-based approach is still a major challenge. In this paper, we present a fast knowledge-based approach (HS-Pharm) that allows the prioritization of cavity atoms that should be targeted for ligand binding, by training machine learning algorithms with atom-based fingerprints of known ligand-binding pockets. The knowledge of hot spots for ligand binding is here used for focusing structure-based pharmacophore models. Three targets of pharmacological interest (neuraminidase, β2 adrenergic receptor, and cyclooxygenase-2) were used to test the evaluated methodology, and the derived structure-based pharmacophores were used in retrospective virtual screening studies. The current study shows that structure-based pharmacophore screening is a powerful technique for the fast identification of potential hits in a chemical library, and that it is a valid alternative to virtual screening by molecular docking.

INTRODUCTION

Structure-based virtual screening of compound libraries is now widely used to prioritize hits for a given protein target of known three-dimensional (3-D) structure. The method of choice in these cases consists of docking numerous compounds in the receptor binding site and scoring the corresponding poses with energy-based scoring functions.[1] Several docking algorithms and scoring functions are available,[2] but the main limitation of this approach resides in the fact that their use is highly target-dependent.[3,4] Even if some general rules have been reported,[4] it is still very difficult to know in advance which combination of docking program and scoring function will give optimal results for a particular target, and as such, it is normal practice to try in consensus a few docking/scoring combinations to identify the most suitable in each case.[5,6]

The use of structure-based pharmacophore models[7–11] for virtual screening is a potential alternative to docking since pharmacophore parametrization is not target-dependent, it may be fuzzy enough to accommodate target flexibility, and it is computationally not expensive. A pharmacophore is the 3-D arrangement of features that are necessary for the binding of small molecules to a macromolecular receptor. In most common applications, pharmacophores are identified from a set of ligands of known activity but unknown receptor structure.[12] Knowledge of the corresponding receptor structure may be incorporated while editing the pharmacophore by generating exclusion spheres mapping protein atoms or, even better, by prioritizing molecular interaction features[13] from protein structures. The major difficulty of the latter approach is the identification of the hot spots for binding, that is, those residues that can form strong interactions with a small-molecular-weight compound. If one or more X-ray structures of protein−ligand complexes are available, then interacting residues can be easily identified, and the corresponding receptor-based pharmacophore models are easier to generate.[9,10]

In many cases, however, especially for the ever-increasing number of proteins solved by structural genomics consortia,[14] only the X-ray structure of the apoprotein is available with no knowledge of the binding mode of putative ligands. One must then only rely on target information to set up reliable pharmacophores. The structure-based pharmacophore (SBP) method[15] implemented in Discovery Studio[16] is a possible approach to this problem. SBP converts LUDI[13] interaction maps within the protein binding site into Catalyst[16] pharmacophoric features: H-bond acceptor, H-bond donor, and hydrophobe. The main limitation of the SBP flowchart is that interaction maps generally consist of hundreds of Catalyst features, which means thousands of possible pharmacophoric hypotheses, and this makes the pharmacophore-based screening of a compound library computationally expensive. In a slightly different approach still conceptually close to SBP, Schüller et al. recently proposed converting the LUDI interaction maps into a virtual ligand described as an alignment-free descriptor vector.[17] In both approaches, the *a priori* knowledge of the most reliable anchoring binding site residues would limit the number of possible pharmacophoric descriptors and considerably speed up the screening process. Targeting the most favored hydrophobic and H-bonding residues of a binding site has been reported in two separate approaches,[18,19] which however still remain incom-

[*] Corresponding author phone: +33-3-90244235, fax: +33-3-90244310, e-mail: didier.rognan@pharma.u-strasbg.fr.
[†] UMR 7175 CNRS-ULP.
[‡] UMR 7551 CNRS.

10.1021/ci800064z CCC: $40.75
Published on Web 06/21/2008

plete with respect to full pharmacophore definition. Grid[20] and SuperStar[21] can be used to identify hot spots by predicting the most favorable positions of atomic probes within the active site. Whereas Grid uses a force field to derive those positions, SuperStar relies on a knowledge-based approach from experimentally determined nonbonded interactions in small molecules' crystal structures.[22] Coupling Grid interaction maps to receptor-based pharmacophore generation was reported several years ago[8] and later refined using an automated protocol.[23] Grid interaction maps have also been applied as a preliminary filter to docking,[24] to detect protein−protein interaction areas,[25] and recently to merge chemical and biological spaces in the FLAP algorithm.[11] Nonetheless, converting interaction maps into pharmacophore queries is still cumbersome and usually requires a significant level of manual intervention, notably the choice of an acceptable energy threshold to derive interaction field minima.

As a conclusion to existing state-of-the art methods, there is still a need for a fast and automated computational approach able to directly convert 3-D atomic structures of protein binding sites into simple and workable pharmacophoric queries. The current study aims at filling this gap by applying machine learning algorithms[26] to identify hot spots for ligand binding in protein binding sites. A 3-D database of protein−ligand structures was first set up and converted into ligand-annotated cavity fingerprints to train several machine learning algorithms to discriminate ligand-interacting from noninteracting protein atoms. A model built using a combination of random forest decision trees was identified as the best to predict interacting atoms in protein binding sites. This model was subsequently used to predict interacting atoms from the known X-ray structure of three proteins of pharmacological interest: neuraminidase, $\beta 2$ adrenergic receptor, and cyclooxygenase-2. Predicted interacting atoms were then utilized as seeds to generate structure-based pharmacophore models, which were further queried to discriminate true ligands from chemically similar decoys. Our approach significantly simplifies pharmacophore generation by restricting the number of interesting anchoring protein atoms and enables the systematic screening of all possible four-feature pharmacophore models to identify potential ligands.

METHODS

Setting up a Data Set of Interacting and Noninteracting PDB Cavity Atoms. The data set used in this study was selected from the third release (2006) of the sc-PDB,[27] which is a collection of druggable ligand-binding sites extracted from the Protein Data Bank.[28] Out of the 4468 sc-PDB entries, several filters were applied to simplify the data set by removing protein−ligand duplicates (at the level of SMILES string for the ligand and E.C. annotation for the protein) and any entry containing cofactors, heme groups, and nonstandard amino acids. The final protein−ligand data set was composed of 3500 entries. For each entry, the binding site was defined from any amino acid for which one atom is located within a 4.5-Å-radius sphere centered on all atoms of the sc-PDB ligand.

The interaction fingerprint (IFP) program[29] was used to identify protein atoms that interact with the corresponding ligand. IFPs are 1-D bit vector representations of the protein−ligand interactions. Eight interaction types were considered for each protein atom: hydrophobic, aromatic (face-to-face), aromatic (edge-to-face), H-bond (protein donor atom), H-bond (protein acceptor atom), ionic (positively charged protein atom), ionic (negatively charged protein atom), and metal complexation. The interactions were identified on the basis of protein and ligand atom types, distances, and angles between atoms. Details of the program have been previously reported elsewhere.[29] Each atom of all 3500 ligand-binding sites was classified as interacting if at least one of the eight bits in the IFP was switched on. The data set contains a total of 623 759 atoms, of which 122 070 are found to be interacting with a ligand (I_A) and 501 689 are found to be non interacting (NI_A), with a ratio NI_A versus I_A of 4.1:1.

Defining Atom-Based Cavity Fingerprints (CFP). Each binding site of the data set was described by use of an atom-based fingerprint which accounts for the properties of the binding site atom a_i, properties of the residue r_i that the atom a_i belongs to, and properties of the environment e_i of atom a_i. Three different fingerprints (CFP1, CFP2, and CFP3) were used, which only differ in the way the residue and environment properties are encoded. Pharmacophoric property assignment (hydrophobe, aromatic, H-bond donor, H-bond acceptor, positive charge, negative charge, and metal) was done on-the-fly using OpenEye's OEChem 1.4.2 library.[30]

Atom Properties. The following pharmacophoric and topological features are registered in a 10-bit vector to describe each atom a_i: hydrophobicity, aromaticity, H-bond donor, H-bond acceptor, positive charge, negative charge, location in the main chain, location in the side chain, and accessibility. With the exception of accessibility, each property is encoded in a binary way to account for the presence (1) or absence (0) of the property itself. The atomic accessibility is defined as follows:

$$\text{accessibility} = \frac{\text{MS}}{4\pi r^2} \times 100 \qquad (1)$$

where MS is the molecular surface calculated with the MS Connolly program[31] using a probe of radius of 1.4 Å, and r is the van der Waals radius of the atom under consideration (default MS values).

Three intervals were selected to encode the accessibility in two bits, on the basis of the computed difference in accessibility between interacting and noninteracting atoms from sc-PDB ligand-binding sites (Supporting Information): 01, accessibility $\leq$ 5%; 11, 5% $<$ accessibility $\leq$ 30%; 10, accessibility $>$ 30%.

Residue Properties. This block depends on the cavity fingerprint definition (CFP1, CFP2, or CFP3).

CFP1 is a nine-bit binary fingerprint accounting for eight physicochemical properties: hydrophobicity, aromaticity, H-bond donor, H-bond acceptor, positive charge, negative charge, metal, and size. The size of a residue is defined according to the number of heavy atoms, and it is encoded in two bits: 10, small (between zero and three heavy atoms); 11, medium (between four and six heavy atoms); 01, large (between 7 and 10 heavy atoms).

CFP2 is an integer fingerprint featuring the number of occurrences of each above-described property in the residue of interest. The size of the residue is encoded in two bits as for CFP1.

CFP3 is a binary fingerprint of 21 bits, where 20 bits account for each of the 20 natural amino acids and one additional bit accounts for metals (Ca, Co, Mg, Mn, and Zn).

Environment Properties. The environment of an atom a_i in the binding pocket is defined by any amino acid for which one atom is located within a 4.5-Å-radius sphere centered on the atom of interest and whose side chain is pointing toward the atom itself. The side chain of a neighboring residue is considered to be pointing toward the atom of interest if the angle formed by the atom, the Cα of the neighboring residue, and the geometric center of the neighboring residue is below 90°. The same encoding as previously reported for the "residue property" block is then collectively applied to all residues selected to contribute to the environment.

The sc-PDB protein cavities are then described by 26 attributes for CFP1 and CFP2 fingerprints and 52 attributes for the CFP3 fingerprint (see exhaustive definition in the Supporting Information).

Classification Models. The data set on which various machine learning algorithms were evaluated consists of 623 759 atoms characterized by an identifier, the corresponding cavity fingerprint (CFP1, CFP2, or CFP3), and a final bit encoding its interacting (I_A) or noninteracting (NI_A) character. Decision trees and naive Bayesian inference were investigated for the classification of interacting and noninteracting atoms in protein binding sites. Decision trees were generated using the Java machine learning workbench WEKA v.3.5.5,[26] while Pipeline Pilot v.6.1[32] was used for naive Bayesian classification. Decision trees work from the top down, selecting at each stage an attribute that best separates the classes. In the current study, the J48 algorithm, which is a Java implementation of Quinlan's C4.5 algorithm,[33] and the random forest approach[34] were used. The main feature of the J48 algorithm is the postpruning of leaves that do not greatly contribute to the predictive accuracy of the tree. It is widely recognized that predictive abilities of trees can be improved by training each tree on a different subset of data, using techniques such as bagging or boosting.[26] For this reason, the AdaBoostM1 algorithm[35] implemented in WEKA was used in conjunction with J48 trees. This algorithm generates bootstrapped samples with replacement from the original data set; at each stage, the accuracy of a tree is used to bias the selection of entries for the next sample, so that poorly predicted entries have a higher probability of being selected, and the next tree can focus on these more difficult cases. Also in random forest, trees are built using samples with replacement from the original data set, and in addition, only a subset of the attributes is used to build each tree. Unlike J48, no pruning is performed in random forest.

A naive Bayesian classifier is based on Bayes' theorem, which relates the conditional and marginal probabilities of two events, A and B, as follows:

$$P(\mathrm{A}|\mathrm{B}) = \frac{P(\mathrm{B}|\mathrm{A})P(\mathrm{A})}{P(\mathrm{B})} \tag{2}$$

where $P(\mathrm{A}|\mathrm{B})$ is the conditional probability of A given B, $P(\mathrm{B}|\mathrm{A})$ is the conditional probability of B given A, $P(\mathrm{A})$ is the prior probability of A, and $P(\mathrm{B})$ is the prior probability of B.

For each object that has to be classified, a naive Bayesian model returns the probability of the object to belong to one class. This classifier assumes events to be equally important and independent. In Pipeline Pilot,[32] a Laplacian correction[36] is used to ensure that attributes that never occur receive a probability value.

Classification models were built using a randomly selected training set corresponding to 25% of the full data set and a test set consisting of the remaining 75% of the data. The criteria used for the evaluation of the performance of each model and for a comparison of different models are sensitivity, specificity, precision(I_A) and precision(NI_A), which are defined as follows:

$$\mathrm{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{3}$$

$$\mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \tag{4}$$

$$\mathrm{Precision}(\mathrm{I_A}) = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{5}$$

$$\mathrm{Precision}(\mathrm{NI_A}) = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}} \tag{6}$$

where TP are true positives, TN true negatives, FP false positives, FN false negatives, I_A interacting atoms, and NI_A noninteracting atoms.

The receiver operating curve (ROC) score[37] was also used for assessing and comparing the performance of the different methods. The ROC score is the area under the ROC curve, which represents the correlation between the false positive rate, defined as 1-specificity, and the sensitivity. The ROC score can vary between 0 and 1, where 1 represents the ideal situation of perfect prediction.

Converting Predicted-Interacting Atoms into Pharmacophore Queries. DiscoveryStudio2.0[16] was utilized as a platform to generate pharmacophore queries from ligand-free protein structures as follows. The "Interaction Generation" protocol was first used to read the input protein PDB file with hydrogen atoms and to output a LUDI interaction map for three features (H-bond donor, H-bond acceptor, and hydrophobic). The latter were subsequently hierarchically clustered according to their type and location, and only cluster centers complementary to the set of previously predicted interacting atoms were kept. The resulting pharmacophores were finally supplemented with 2-Å-radius exclusion spheres located at the Cα atomic coordinates of the user-defined binding site residues. In cases where these exclusion spheres were imperfectly mimicking the protein surface, additional spheres were added, notably for bulky side chains (e.g., Arg and Trp) folding inward the binding site center.

Compound Library Setup and Screening. For each of the three investigated targets, a compound library was customized to contain as many chemically diverse known actives and decoys describing a similar chemical space. True actives were retrieved from the literature (see the Supporting Information). The structures of the targeted ligands in a 2-D sd file were ionized at physiological pH with Filter v.2.0.1[38] and standardized with JChem v.3.2.3,[39] and 3-D coordinates were generated with Corina v.3.4.[40]

3-D coordinates were saved either in sd file format for a pharmacophore search or in mol2 file format for docking. Decoys were taken from our in-house repository of commercially available druggable ligands[41] and selected to span global property ranges (molecular weight, H-bond donor count, H-bond acceptor count, logP, and negative and positive charge counts) similar to those of the set of actives. To reduce the risk that potentially active compounds might be selected by chance, a maximal Tanimoto similarity threshold of 0.3 computed from SciTegic ECFP4 circular fingerprints[32] was applied to remove decoys close to any of the true actives.

A maximum of 250 conformers were generated in DS Catalyst for each compound using the FAST parameter settings with an upper energy threshold of 20 kcal mol^{-1}. The "Screen Library" protocol in DS was used to screen the compound libraries against all possible combinations of three-, four-, and five-feature pharmacophore models, using a minimum interfeature distance of 0.5 Å and rigid fitting. Complete screening timings ranged from 15 min to 3 h on a 2.4 GHz PC running Windows XP with 1.25 GB of RAM.

Docking. Standard parameters of FlexX2.2,[42] Surflex2.11,[43] and two times speed-up settings of Gold3.2[44] were used to dock the above-described libraries. The binding site was defined in FlexX (receptor description file), Surflex (protomol file), and Gold (gold.conf file) from the 3-D coordinates of binding cavities, as predicted by MOE Site Finder.[45] 3-D coordinates of the targets (2qwf, 2rh1, and 5cox) were directly taken from the sc-PDB database[46] without any further energy refinement. Whereas FlexX used a PDB file format without explicit definition of hydrogen atom positions, Surflex and Gold utilized a mol2 file format with explicit hydrogen atoms. The top-ranked pose, according to the native scoring function of each docking program (FlexXscore, $-\log(K_d)$, Goldscore) was saved for further postprocessing with in-house Perl and shell scripts to rank compounds by decreasing docking scores.

RESULTS AND DISCUSSION

The basic idea behind the herein-proposed HS-Pharm flowchart (Figure 1) is to use machine learning algorithms to classify protein cavity atoms as noninteracting or interacting and exclusively focus the receptor-based pharmacophore definition on the latter atoms. To achieve this goal, the first step is to develop a database of atoms lining ligand-binding sites, annotate them according to their anchoring potential (interacting, noninteracting), and describe them by a fingerprint with some physiochemical relevance.

The Data Set. As a source of ligand-binding sites, we chose the sc-PDB data set,[27] which presents several advantages: (1) the corresponding ligand is characterized from a pharmacological and not a structural point of view, (2) ligand-binding sites are filtered by ligand-based and structure-based topological filters to retain druggable cavities only, (3) all binding sites are carefully annotated at the biochemical level. Starting from 4468 sc-PDB entries, 3500 binding cavities were finally selected and randomly separated into a training set (25% of entries) and a test set (75% of entries). Analysis of the biochemical annotation (at the E.C. annotation level) of all entries in both sets suggests that there is no major bias in the splitting procedure (Figure 2A) and that the E.C. annotation of enzymes in both sets mirrors that of the Protein Data Bank[28] for enzymes. As already noticed for the full sc-PDB,[27] the data set is enriched in enzymes (ca. 70% of entries) with respect to the PDB, which simply corresponds to the higher propensity of enzymes to be cocrystallized with a druglike ligand.

For each complex of the data set, the corresponding protein−ligand molecular interactions for each cavity atom were computed with the IFP program.[29] Out of the 623 759 protein cavity atoms stored in our data set, about one-fourth (122 070) were found to be interacting with a ligand. The distribution of molecular interaction types is again very similar in both the test set and the training set (Figure 2B). The prevalence of hydrophobic contacts among registered interactions in IFP bit strings (ca. 70%) reflects the abundance of carbon atoms in PDB complexes and shape recognition between a protein cavity and its cognate ligand. Whereas aromatic contacts are equally distributed among edge-to-face and face-to-face interactions (3% for each), H bonds are significantly more frequent from protein donor atoms (13%) than from protein acceptor atoms (7%). The same discrepancy is observed by comparing the occurrence of salt bridges with protein cationic atoms (2% of all interactions) and anionic atoms (1%). These discrepancies only reflect a bias in the chemical space described by PDB ligands. Hence, among the 5525 currently registered sc-PDB ligands, there are nearly twice more anionic ligands (e.g., nucleotide analogues) than cationic compounds.

Assessment of Classification Models. After tagging protein cavity atoms as ligand-interacting and noninteracting, we next need to encode topological and physicochemical properties of protein cavity atoms into a simple fingerprint from which machine learning algorithms may learn to distinguish interacting from noninteracting atoms. Three slightly different CFPs and classification algorithms (naive Bayes, decision trees, and random forest) were compared for this purpose. A total of 25% of the data set was used for learning, and the remaining 75% was utilized for testing. Default values in WEKA were used in building the decision trees; for random forest, the increase in the number of trees (default is 10) was investigated, but no improvement in predictions was observed. For each CFP, random forest and boosting combined with the J48 algorithm yield similar predictions (Table 1). All CFPs are relatively similar in specificity, precision, and ROC score, but the sensitivity is lower for CFP1 compared to those for CFP2 and CFP3. This is likely to be due to the nature of the fingerprints, since CFP1 collectively describes residues and the environment by property type, whereas CFP2 and CFP3 allow a better description of individual cavity residues and thus of corresponding individual atoms. Mixing integers with bits in the fingerprint (CFP2) does not affect the accuracy of the corresponding models with respect to a full bit string definition (CFP1 and CFP3). If we now compare the classification performance of the naive Bayesian classifier to the decision trees, it can be noticed that a higher sensitivity is achieved by all CFPs, but at the cost of a significant decrease in precision(I_A) and in specificity, which means that, while more interacting atoms are correctly classified as interacting, more noninteracting atoms are misclassified as interacting. CFP3 achieved the highest sensitivity and lowest specificity, while CFP2 has the lowest sensitivity and highest

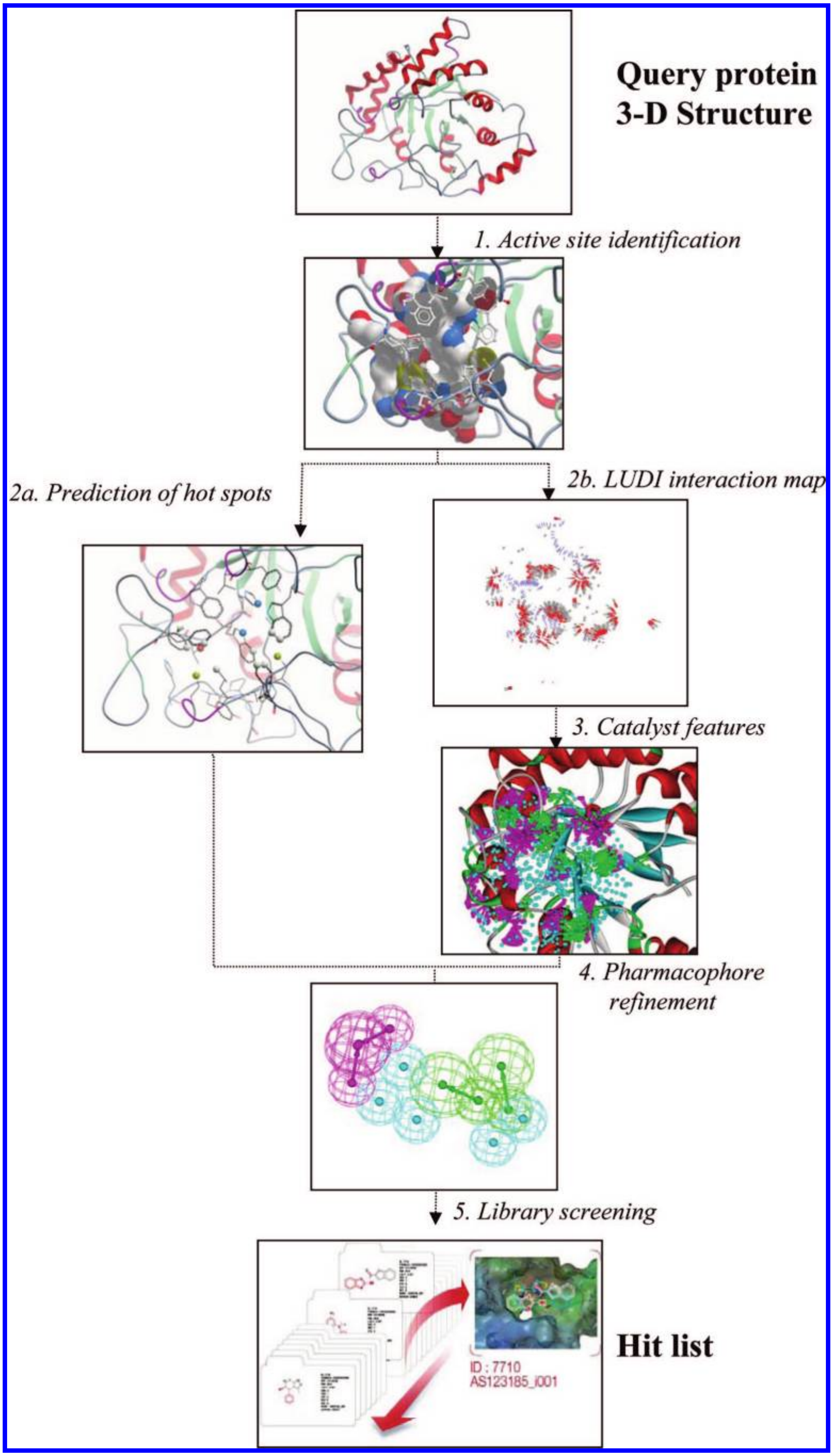

Figure 1. HS-Pharm flowchart. Starting from a 3-D structure of a target, the most relevant binding site is selected manually or automatically (step 1), and it is used to prioritize the most interesting interacting atoms (step 2a) and to generate an interaction map with few probe atoms (step 2b). The interaction map is converted into Catalyst pharmacophoric features (step 3) and further simplified by focusing on cavity atoms predicted to be interacting with a putative ligand (step 4). The simplified features are transformed into all possible three-feature, four-feature, and five-feature pharmacophores, which are sequentially screened (Step 5) to find putative hits.

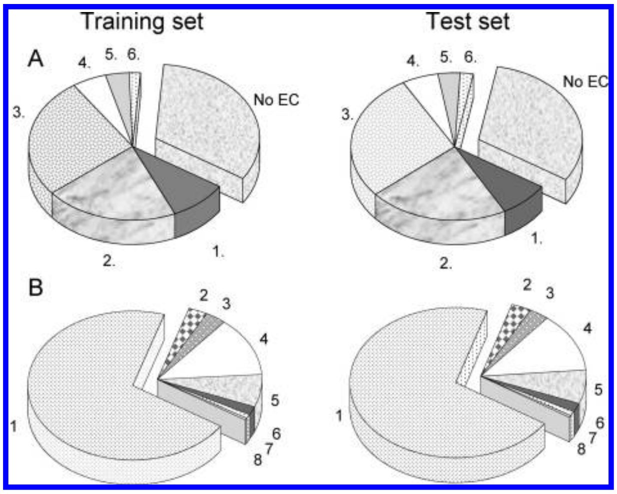

Figure 2. Properties of the test set and training set (sc-PDB protein−ligand complexes) used for classifying protein atoms. (A) Distribution of sc-PDB entries according to the E.C. classification (1, oxidoreductases; 2, transferases; 3, hydrolases; 4, lyases; 5, isomerases; 6, ligases; No EC, no E.C. number attributed). (B) Molecular interactions encoded in all atom-based molecular interaction fingerprints (1, hydrophobic; 2, edge-to-face aromatic; 3, face-to-face aromatic; 4, H-bond (protein is donor); 5, H-bond (protein is acceptor); 6, ionic interaction (protein is positively charged); 7, ionic interaction (protein is negatively charged); 8, metal complexation.

Table 1. Evaluation of Classification Models on the Test Set

model	CFP	Se[a]	Sp[b]	Pr (I_A)[c]	Pr (NI_A)[d]	ROC[e]
random forest	CFP1	0.27	0.96	0.60	0.84	0.82
	CFP2	0.37	0.94	0.61	0.86	0.81
	CFP3	0.37	0.94	0.61	0.85	0.82
AdaBoostM1(J48)	CFP1	0.27	0.96	0.60	0.84	0.82
	CFP2	0.36	0.94	0.62	0.86	0.82
	CFP3	0.37	0.94	0.61	0.85	0.82
naive Bayes	CFP1	0.76	0.59	0.31	0.91	0.74
	CFP2	0.72	0.64	0.33	0.90	0.73
	CFP3	0.80	0.56	0.30	0.92	0.75

[a] Sensitivity. [b] Specificity. [c] Precision (interacting atoms). [d] Precision (noninteracting atoms). [e] Area under the ROC curve.

Table 2. Evaluation of Additional Classification Models on the Test Set

model	CFP	Se[a]	Sp[b]	Pr (I_A)[c]	Pr (NI_A)[d]	ROC[e]
RF/ThresholdSelector[f]	CFP1	0.68	0.79	0.44	0.91	0.82
	CFP2	0.66	0.81	0.46	0.91	0.81
	CFP3	0.66	0.81	0.46	0.91	0.81
Vote[g]	CFP1	0.41	0.91	0.53	0.86	0.78
	CFP2	0.48	0.90	0.48	0.90	0.81
	CFP3	n.a.[h]	n.a.	n.a.	n.a.	n.a.

[a] Sensitivity. [b] Specificity. [c] Precision (interacting atoms). [d] Precision (noninteracting atoms). [e] Area under the ROC curve. [f] Random forest with threshold selection (see text). [g] Voting algorithm combining random forest alone and random forest with threshold selection. [h] Not available.

specificity. This is most likely due to the fact that, while CFP1 and CFP3 are binary fingerprints, CFP2 is a mixed fingerprint, which combines a binary part to an integer part; this probably affects the naive Bayesian Pipeline Pilot classification.

We thought that specificity was an important descriptor of the model since we wished to define a limited set of potentially interacting atoms to simplify the corresponding pharmacophores. We therefore prioritized random forest and decision trees to further improve the limited sensitivity while keeping the specificity of our preliminary models. As previously reported, interacting atoms are less abundant in the data set than noninteracting atoms. Several methodologies can be used to improve predictions on the least represented class in an unbalanced data set. One such methodology implemented in WEKA is the meta learner algorithm ThresholdSelector,[26] which selects a probability threshold on the classifier's output in order to optimize the *F* measure on the least represented class. The *F* measure is defined as follows:

$$F = \frac{2TP}{2TP + FP + FN} \tag{7}$$

where TP are true positives, FP are false positives, and FN are false negatives.

When compared to a simple random forest, the use of this meta learner increases the sensitivity, as expected, but decreases the specificity and the precision (I_A) for all three fingerprint types (Table 2). We thus tried a voting algorithm combining random forest alone and random forest with threshold selection to average probability estimates for interacting and noninteracting atoms. The voting algorithm could not be used in combination with the CFP3 fingerprint due to memory limitations in WEKA. However, for the other two cavity fingerprints, the combined model provides a good compromise between the two separate initial classifications. CFP2 descriptors lead to a slightly higher ROC score and sensitivity than CFP1. In addition, the corresponding model reaches a better balance between sensitivity and precision. The specificity is still good enough to limit the number of predicted interacting atoms. For this reason, the CFP2 cavity fingerprint was further selected for retrospective pharmacophore elucidation for three targets (see below).

Evaluation on an External Validation Set. To further probe the predictive ability of the above classification model, an additional external validation set was created from entries recently added to the fourth release of the sc-PDB database. Of the 2057 new entries found, only those whose E.C. number was absent in the original data set of 3500 entries were selected. The external validation set was composed of 114 new entries, on which none of the previous classifiers had been trained. In addition to the previous definition of binding sites used in the original data set (see Methods), larger cavities were created by selecting residues within 6.5 Å of the ligand. Last, for 10 randomly selected entries in the external validation set, the binding site was detected on-the-fly using the MOE cavity detection algorithm.[45] The Site Finder module in MOE identifies all pockets in a given protein apo structure and ranks them as possible binding sites. For each entry, the cavity identified by MOE which was closest to the true binding site (according to the ligand-based definition) was extracted and used for predictions. This cavity definition truly reproduces the situation in which the methodology described here would be used to prioritize hot spots from apoprotein structures.

As can be seen from Table 3, predictions obtained using the previously described voting approach are still in good agreement with previous predictions on the test set, whatever

Table 3. Atom-Based Predictions on the Validation Set

binding site definition	Se[a]	Sp[b]	Pr (I_A)[c]	Pr (NI_A)[d]	ROC[e]
4.5-Å-radius sphere	0.42	0.90	0.48	0.87	0.77
6.5-Å-radius sphere	0.46	0.87	0.30	0.93	0.76
MOE SiteFinder	0.47	0.89	0.44	0.91	0.76

[a] Sensitivity. [b] Specificity. [c] Precision (interacting atoms). [d] Precision (noninteracting atoms). [e] Area under the ROC curve.

the binding site definition. We can thus assume that this classification model has been trained on a training set which sufficiently samples known cavity atoms in order to derive relevant predictions. In addition, the use of an automated procedure to detect the ligand-binding cavity does not alter the accuracy of the derived model either. We can thus conclude that the voting model is robust enough to be efficiently used for predicting ligand-anchoring atoms in protein binding sites.

Defining and Querying Structure-Based Pharmacophore Models. As discussed in the introduction, the knowledge of hot spots in protein cavities can be applied to the generation of structure-based pharmacophore models, which can be further queried to discover appropriate ligands by virtual screening. Three proteins, which were not present in the training set used to train the classification models, were selected to test the methodology. The Site Finder module in MOE was used to extract the binding sites from the proteins, as described above, and CFP2 cavity descriptors were calculated for each atom of the corresponding pockets. The combined model (random forest and random forest with ThresholdSelector) was used for predicting interacting atoms in the binding sites.

Neuraminidase (NA). The influenza virus neuraminidase is a tetrameric glycoprotein target for anti-infuenza drugs.[47] This target was chosen for the highly polar nature of its sialic acid binding site, which is formed of 28 residues with a high abundance of charged amino acids. Our classification model applied to the ligand-depleted 2qwf PDB entry[48] predicts 32 atoms to be of interest for interacting with a potential NA inhibitor (Figure 3A). Out of these 32 atoms, 14 are less than 4.5 Å away from the 2qwf ligand, whose coordinates were not taken into account in the classification. A pharmacophore was derived from the LUDI interaction map, and only those features close to any of the 32 atoms of interest were selected, thus limiting the number of pharmacophoric features to seven: one H-bond acceptor toward Lys292 (Acceptor1); one H-bond acceptor toward Arg118 and Arg371 (Acceptor 2); three H-bond donors toward Glu277 (Donor 1), Glu119 (Donor2), and Glu276 (Donor3); one hydrophobic feature close to Ile222 and Arg224 side chains (Hydrophobe1); and one hydrophobic feature close to Arg152, Trp178, and Ser179 side chains (Hydrophobe2). A total of 37 exclusion spheres were added to the pharmacophore model, which is shown in Figure 3B.

A compound library was set up from eight known NA inhibitors (see the Supporting Information) and 792 decoys randomly selected from an in-house database of commercially available druglike compounds (Bioinfo database).[41] In order to avoid biasing *in silico* screening results, decoys were carefully selected to span the same chemical space as true NA inhibitors. All possible three-feature, four-feature, and five-feature pharmacophores were serially screened with DS Catalyst for their propensity to retrieve true NA inhibitors. Using three-feature pharmacophores (P3 model, Figure 4D), all eight active compounds were retrieved, but the majority of the decoys also passed the pharmacophores (sensitivity = 1.0; specificity = 0.1). With the four-feature pharmacophores (P4 model), six out of eight actives were retrieved, and the number of false positives considerably decreased (sensitivity = 0.75; specificity = 0.68). Finally, using five-feature pharmacophores (P5 model), only one active was retrieved, and the number of decoys retrieved decreased even more (sensitivity = 0.25; specificity = 0.98). Screening with four-feature pharmacophores thus represents the best compromise, as screening with three-feature pharmacophores is too permissive, while screening with five-feature pharmacophores is too restrictive. We then analyzed the results of this *in silico* screening to check whether some features are prevalently matched by true positives. Heat maps, which highlight the matched and unmatched pharmacophoric features, were generated for all positives, which means any conformer passing any of the four-feature pharmacophores. It can be noticed that the majority of true positives (conformers of true ligands) match Donor1 and Donor2, which is not the case for positives (Figure 4A,B). To automatically select those features that are preferentially matched by true positives, we computed the matching frequency of each feature for all conformers passing the pharmacophores and compared the corresponding frequencies for true positives and positives (Figure 4C). The differential frequency (DF) in matching frequency F for a particular feature f between true positives (TP) and positives (P) was computed as

$$\mathrm{DF} = F_f(\mathrm{TP}) - F_f(\mathrm{P}) \qquad (8)$$

where $F_f = n/N$, n being the number of matches and N the number of conformers fulfilling the pharmacophore query.

As previously noticed from the heat maps, two features (Donor1 and Donor2) are more frequently matched by true positives than by positives (DF > 0.15). We thus studied the influence of constraining these features in postprocessing the hit list. It is important to recall at this point that only information about true inhibitors (as it is commonplace for pharmacophore elucidation) and not their binding mode is used to postprocess screening results. When entries that did not match Donor1 were discarded, the same sensitivity was maintained (0.75), but the specificity considerably increased to 0.89 (P41 model, Figure 4D). When entries that did not match Donor2 were discarded (P42 model, Figure 4D), the sensitivity decreased to 0.50 and the specificity increased to 0.90. If entries that did not match both donors were discarded (P43 model, Figure 4D), the sensitivity decreased to 0.37, while the specificity increased to 0.95. We can then conclude that the proposed receptor-based pharmacophore, with two H-bond acceptors, three H-bond donors, and two hydrophobic groups can be used in library screening to find new potential NA inhibitors using all possible combinations of four features, while constraining the Donor1 feature to be necessarily fulfilled.

β2 Adrenergic Receptor (ADRB2). The *β*2 adrenergic receptor is a G-protein coupled receptor (GPCR) that binds adrenaline and noradrenaline to regulate cardiovascular and pulmonary functions.[49] Its binding site can be considered a prototypical druggable binding site with a good balance of polar and hydrophobic residues. The crystal structure of

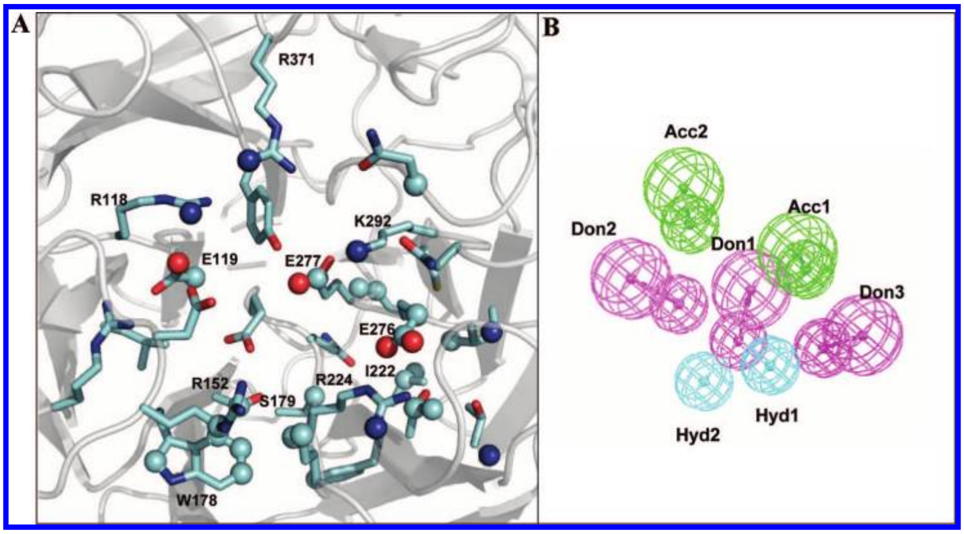

Figure 3. Neuraminidase structure-based pharmacophore. (A) Predicted interacting atoms (spheres) in the 2qwf ligand binding site. (B) Pharmacophoric features: H-bond acceptor features are colored in green; H-bond donor features are colored in magenta; hydrophobic features are colored in blue. Exclusion spheres are not displayed for sake of clarity.

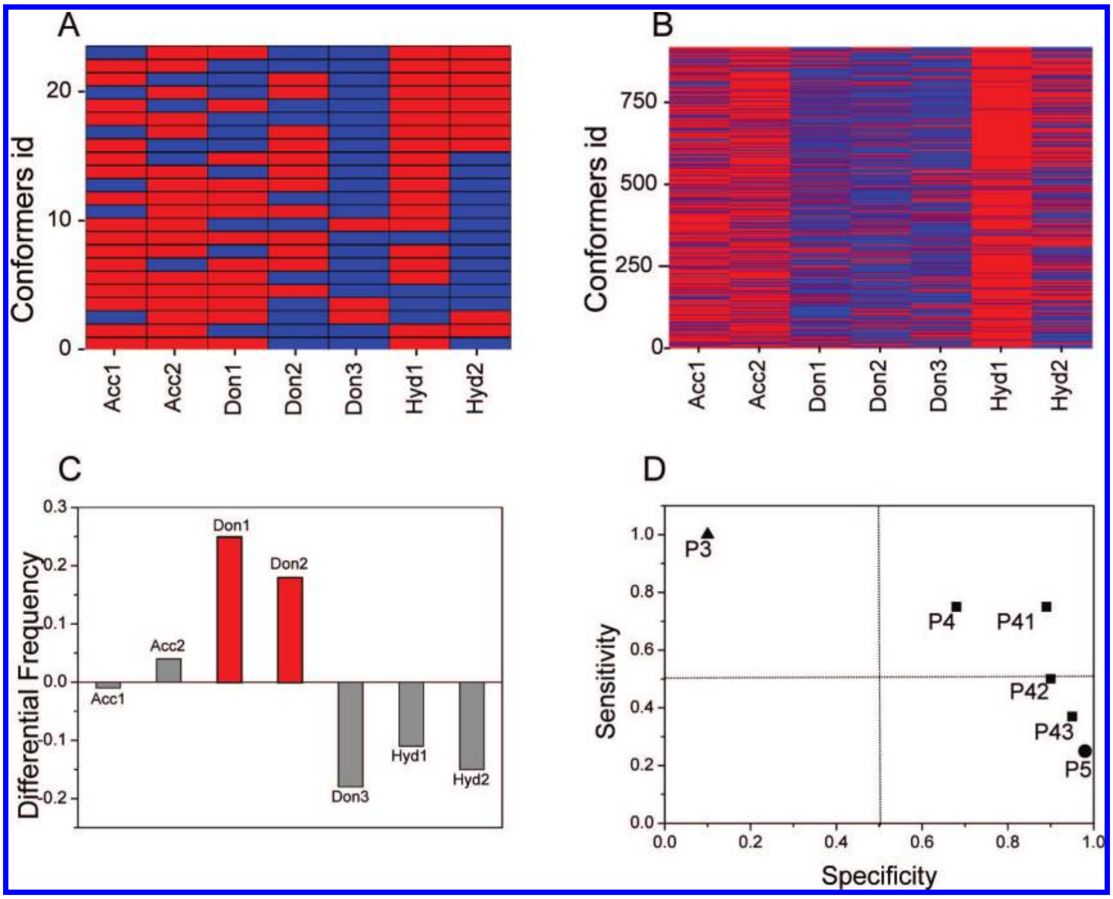

Figure 4. Heat plots for neuraminidase inhibitor screening with 4-feature pharmacophore models. Matched features are colored in red; unmatched features are colored in blue. (A) Actives retrieved in the hit list. (B) Full hit list. (C) Differential occurrence in matching frequency between true positives and positives. (D) Sensitivity vs specificity of various pharmacophore searches: P3, three-feature pharmacophores; P4, four-feature pharmacophores; P41, four-feature pharmacophores with donor 1 fixed; P42, four-feature pharmacophores with donor 2 fixed; P43, four-feature pharmacophores with donors 1 and 2 fixed; P5, five-feature pharmacophores.

ADRB2 in complex with the inverse agonist carazolol was recently published.[50] The ligand-binding site of this receptor (2rh1 entry), is formed by 44 residues (Figure 5A). From the 67 atoms which were predicted to be interacting with a putative ligand, a pharmacophore model with two H-bond acceptor features, three H-bond donor features, and five hydrophobic features was identified (Figure 5B): one H-bond acceptor toward Asn293 (Acceptor1); one H-bond acceptor toward Asn312 (Acceptor2); two H-bond donors toward the carboxylic acid moiety of Asp113 (Donor1 and Donor2); one H-bond donor toward Ser203 (Donor3); one hydrophobic feature close to Val114 and Phe290 side chains (Hydrophobe1); one hydrophobic feature close to three aromatic side chains (Phe193, Phe289, and Tyr308; Hydrophobe2);

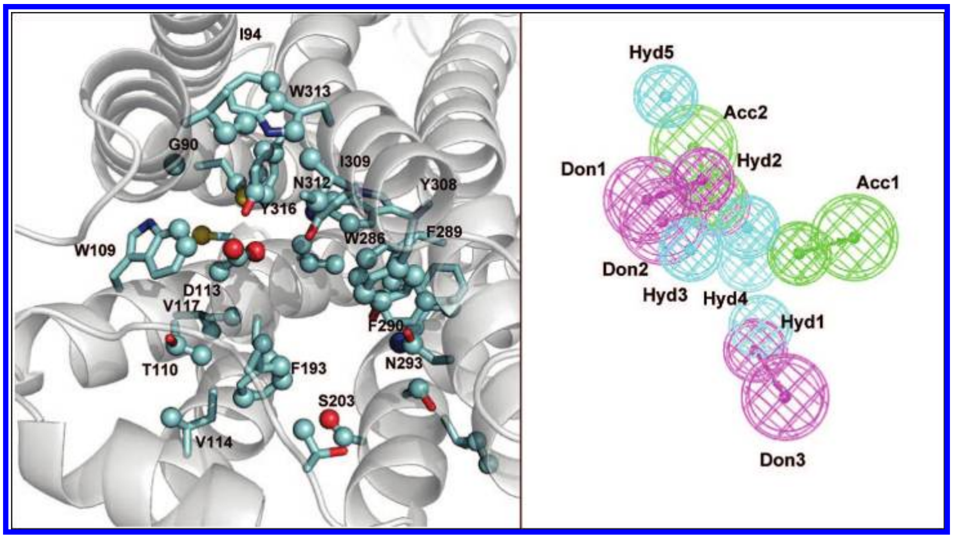

Figure 5. $\beta 2$ adrenergic receptor structure-based pharmacophore. (A) Predicted interacting atoms (spheres) in the 2rh1 ligand binding site. (B) Pharmacophoric features: H-bond acceptor features are colored in green; H-bond donor features are colored in magenta; hydrophobic features are colored in blue. Exclusion spheres are not displayed for sake of clarity.

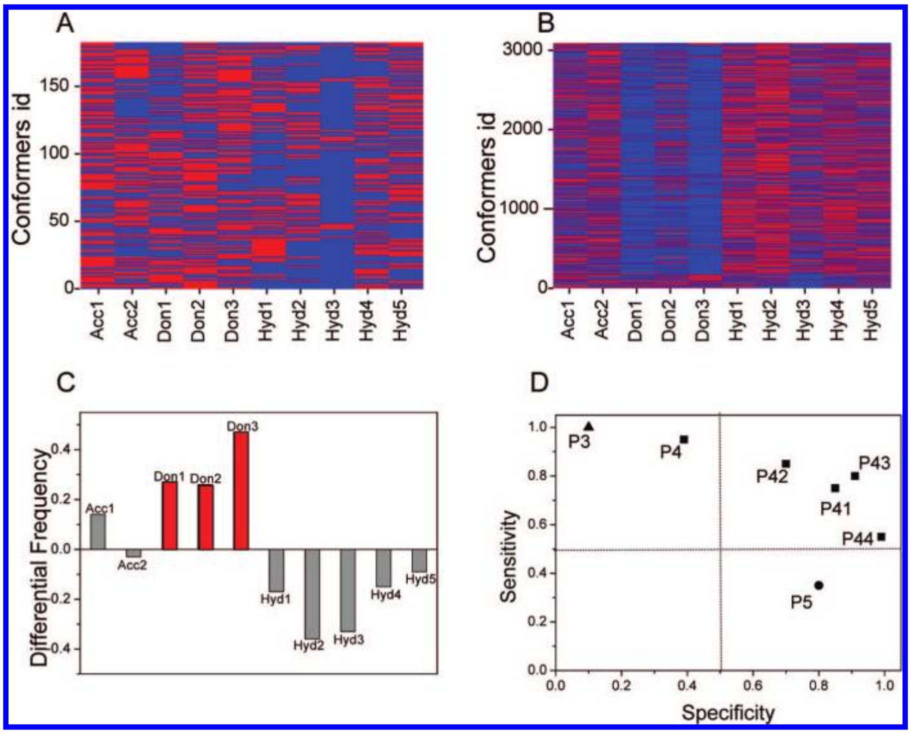

Figure 6. Heat plots for $\beta 2$ adrenergic receptor ligand screening with four-feature pharmacophore models. Matched features are colored in red; unmatched features are colored in blue. (A) True positives. (B) Positives. (C) Differential feature matching frequencies between true positives and positives. (D) Sensitivity vs specificity of various pharmacophore searches: P3, three-feature pharmacophores; P4, four-feature pharmacophores; P41, four-feature pharmacophores with donor 1 fixed; P42, four-feature pharmacophores with donor 2 fixed; P43, four-feature pharmacophores with donor 3 fixed; P44, four-feature pharmacophores with donors 1 and 3 fixed; P5, five-feature pharmacophores.

one hydrophobic feature close to Thr110 and Phe193 side chains (Hydrophobe3); one hydrophobic feature close to Val117, Trp286, and Phe290 side chains (Hydrophobe4); and one hydrophobic feature close to Gly90, Ile94, Trp109, Ile309, Trp313, and Tyr316 (Hydrophobe5). A total of 77 exclusion spheres were added to the final pharmacophoric model, which is shown in Figure 5B.

A $\beta 2$ receptor-targeted compound library was set up from 20 known active compounds (see the Supporting Information) and 980 chemically similar decoys. To check whether the cavity-based pharmacophore may be fuzzy enough to accommodate ligands with different functional effects, we explicitly selected as true actives six inverse agonists, five partial agonists, four antagonists, and five full agonists of the $\beta 2$ receptor. Again, all possible combinations of three-feature, four-feature, and five-feature pharmacophores were evaluated for their ability to selectively retrieve known actives. Screening with three-feature pharmacophores (P3 model, Figure 6D) was far too permissive since it returned 898 hits, among which were all 20 active compounds (sensitivity = 1.0; specificity = 0.10). Screening with four-feature pharmacophores (P4 model) returned fewer compounds (616 hits), among which were 19 active compounds (sensitivity = 0.95; specificity = 0.39). However, its

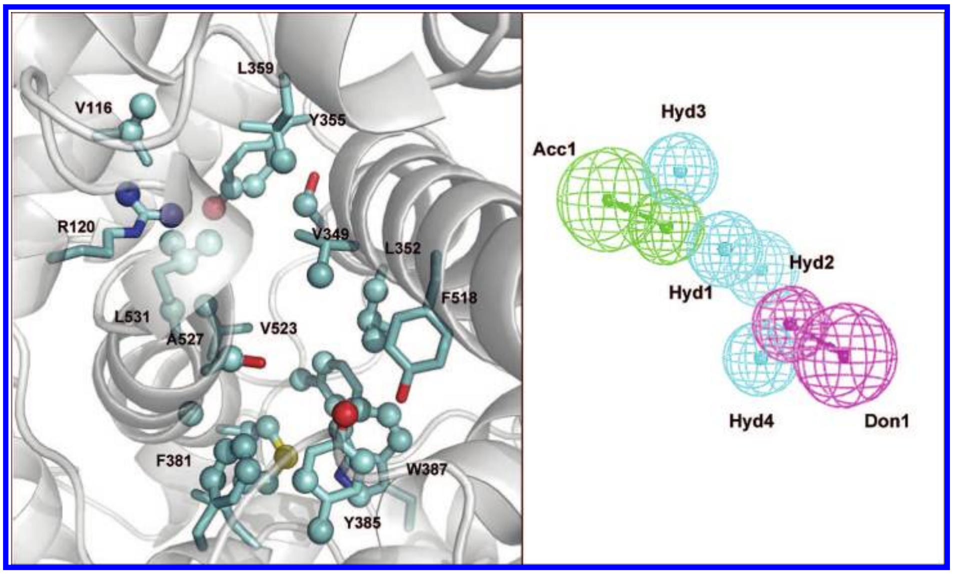

Figure 7. Cyclooxygenase-2 structure-based pharmacophore. (A) Predicted interacting atoms (spheres) in the 5cox ligand binding site. (B) Pharmacophoric features: H-bond acceptor features are colored in green; H-bond donor features are colored in magenta; hydrophobic features are colored in blue. Exclusion spheres are not displayed for sake of clarity.

specificity is still too low to be of real use in prioritizing hits. Finally, screening with five-feature pharmacophores (P5 model) returned 200 hits, seven of which were active compounds (sensitivity = 0.35; specificity = 0.80). Again, the four-feature pharmacophores achieve the better balance between specificity and sensitivity. Looking at the difference in feature-matching frequencies between true positives and positives (Figure 6C), the three H-bond donor features appear to be extremely important for true active identification and were thus iteratively fixed in postprocessing the hit list. The best results are obtained by screening with four-feature pharmacophores with either Donor1 (P41 model) or Donor3 (P43 model) fixed. If both donors are kept fixed (P44 model), there is a loss in sensitivity (0.55), but the specificity is highly increased (0.99). Depending on the basic aim of the virtual screening, two different queries might be set up. In order to retrieve chemically diverse actives, a single constraint on either Donor1 or Donor 3 will achieve the best possible sensitivity but at the cost of a bigger hit list. Alternatively, optimizing the hit rate within the shortest possible hit list will be obtained by constraining both Donors 1 and 3 in a four-feature pharmacophore query.

Interestingly, none of these pharmacophore models were able to distinguish between agonists, antagonists, partial agonists, and inverse agonists. Although binding of either full agonists or partial agonists is known to be accompanied by conformational changes at the transmembrane binding cavity of the $\beta2$ receptor,[51] high-resolution X-ray structures of inactivated and photoactivated meta-II states of bovine rhodopsin are surprisingly similar when looking at the transmembrane helical bundle.[52] Moreover, we recently identified full agonists of the CCR5 chemokine receptor using a structure-based approach applied to the presumed inactive state model of the latter receptor,[41] thus illustrating the subtle differences occurring at the level of GPCR−ligand interactions when comparing ligands with different functional outcomes. Interestingly, the receptor-based model (either P41 or P43 model) was suitable in discarding ligands binding to other adrenergic receptor subtypes ($\beta3$ and α1a subtypes) with, as expected, a specificity closely related to the distance between the target and the reference binding sites.[53] For highly related binding cavities (e.g., $\beta2$ and $\beta3$ ligand-binding sites: 23/30 conserved residues), the ligands are too similar, and the reference structure-based pharmacophore hardly discriminates $\beta2$ from $\beta3$ ligands (three out of five selective $\beta3$ receptor ligands were still selected by the $\beta2$-based pharmacophore, data not shown). For a more divergent cavity (e.g., the adrenergic α1a cavity: 17/30 conserved residues with the $\beta2$ binding site), the corresponding ligands are sufficiently different to be discarded by a structure-based pharmacophore (only five out of 50 chemically diverse α1a ligands selected, data not shown).

Cyclooxygenase-2 (COX2). Cyclooxygenase-2 is an enzyme biosynthesized during the inflammation process which catalyzes the conversion of arachidonic acid into prostaglandin H,[54] and it is a major target for nonsteroidal anti-infammatory drugs.[55] The COX2 target is interesting since its ligand-binding site exhibits ligand-induced rearrangements of some key residues (Arg 120, Tyr355, and Val523), which prevent the cross-docking of COX2 inhibitors.[56] The ligand-binding site, as identified with MOE from the apoprotein structure (5cox PDB entry), is formed by 19 residues and presents a strong hydrophobic nature (Figure 7A). From the 45 atoms which were predicted to be interacting, a pharmacophore model with one H-bond acceptor, one H-bond donor, and four hydrophobic features was generated (Figure 7B): one H-bond acceptor toward Arg120 (Acceptor1); one H-bond donor toward the hydroxyl group of Tyr385 (Donor1); one hydrophobic feature close to Val349, Ala527, and Leu531 side chains (Hydrophobe1); one hydrophobic feature close to Leu352, Val523, and Ala527 side chains (Hydrophobe2); one hydrophobic feature close to Val116, Val349, Tyr355, Leu359, and Leu531 side chains (Hydrophobe3); and one hydrophobic feature close to Phe381, Tyr385, Trp387, and Phe518 (Hydrophobe4). A total of 39 exclusion spheres were added to the final pharmacophore model (Figure 7B).

A COX-2 targeted molecular database was created from 10 known inhibitors (see the Supporting Information) and 990 chemically similar decoys selected from the Bioinfo

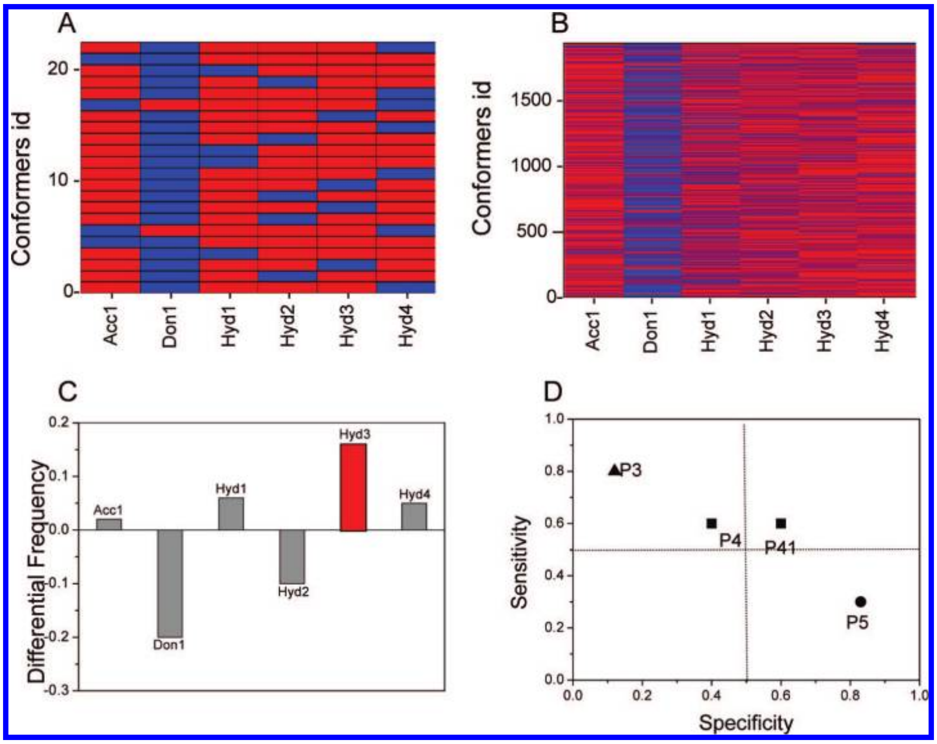

Figure 8. Heat plots for screening cyclooxygenase-2 inhibitors with four-feature pharmacophore models. Matched features are colored in red; unmatched features are colored in blue. (A) True positives. (B) Positives. (C) Differential feature matching frequencies between true positives and positives. (D) Sensitivity vs specificity of various pharmacophore searches: P3, three-feature pharmacophores; P4, four-feature pharmacophores; P41, four-feature pharmacophores with hydrophobe 3 fixed; P5, five-feature pharmacophores.

database. On one hand, screening the compound library against all combinations of three-feature and four-feature pharmacophores (P3 and P4 models, Figure 8D) was too permissive and returned 868 and 590 hits, respectively, with a very modest sensitivity for the four-feature pharmacophore combinations (Figure 8). On the other hand, screening with five-feature pharmacophores (P5 model) was more specific and returned 164 hits, among which were three actives (sensitivity = 0.3; specificity = 0.83). A single difference in the frequency of matched features between true positives and positives can be observed for feature Hydrophobe 3 (Figure 8). Fixing this feature in postprocessing the hit list obtained from a four-feature-based pharmacophore search (P41 model) does not improve sensitivity (Se = 0.6) and only marginally enhances the specificity (Sp = 0.6) of the model.

HS-Pharm Search vs Docking. We believe it would not be appropriate to compare the accuracy of our HS-Pharm searches with that of existing ligand-based or protein−ligand based pharmacophores, since all reported pharmacophores[57–66] for the three targets were queried with different compound libraries, out of which the decoys were usually randomly selected from druglike compounds without assessing that true actives and decoys really span the same chemical space. We can thus assume, in agreement with a recent study on the importance of decoy selection,[67] that the screening accuracy of reported pharmacophore models is in most cases largely overestimated. The most reliable comparison to our viewpoint consists in comparing our method with molecular docking, since both approaches rely on the 3-D coordinates of the target protein and on an identical compound library. For the three targets under investigation, the same targeted-focused libraries as those used above were thus docked into their respective binding sites. To avoid target dependency,[4] three of the best-performing docking tools (Gold, FlexX, and Surflex) in combination with their native scoring function, were then used to rank screened compounds by decreasing docking scores. For two of the three test cases (NA and ADRB2), the HS-Pharm approach outperformed the best possible docking tool in a screening scenario where the top-ranked compounds (ca. 1%) would be experimentally tested for *in vitro* binding (Figure 9). Due to the higher number of less-specific hydrophobic features, the COX-2 structure-based pharmacophores are still too fuzzy to be really informative and therefore less competitive than a pure docking approach (Figure 9). Even if multiple pharmacophores are serially screened in the HS-Pharm approach, it is still much faster than docking (CPU timings for docking were ca. 10−12 h vs 5−20 min for four-feature HS-Pharm screening), and therefore it is a true alternative to docking in most cases.

Advantages and Drawbacks of the HS-Pharm Approach. The worldwide development of structural genomics programs enables the identification of high-resolution 3-D structures of an ever increasing number of potentially interesting targets for which small-molecular-weight ligands need to be identified.[68] Searching compound libraries for potential hits that satisfy receptor-based pharmacophores is one of the computational methods of choice in this case, since molecular docking is computationally expensive, and it is supposedly relatively inaccurate to identify virtual hits from apoprotein structures.[69] The SBP approach[15] is particularly interesting in such scenarios, as it directly generates ready-to-use pharmacophore queries from the ligand-free protein 3-D atomic coordinates. However, a major drawback of the current method is the combinatorial explosion of possible pharmacophoric queries with the increase of selected features. For two of the three targets investigated in the current study, the number of possible pharmacophores is far too high to

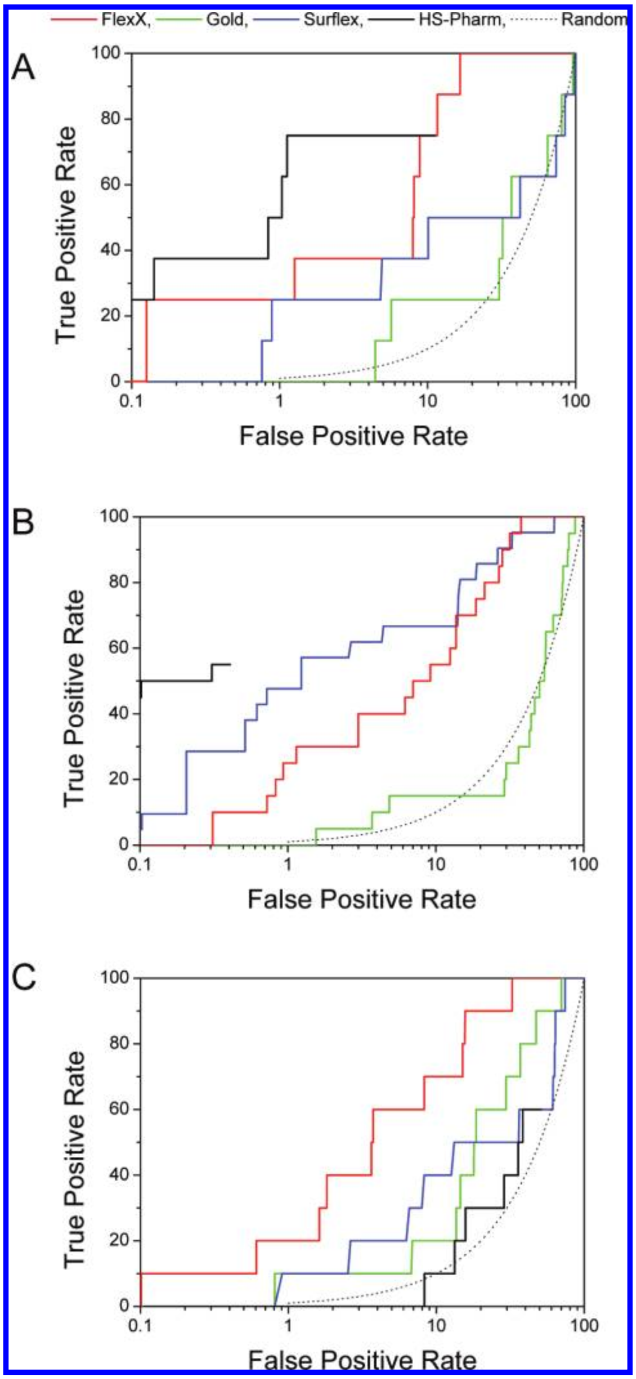

Figure 9. ROC plots from structure-based screening of three targets (A, neuraminidase; B, $\beta2$ adrenergic receptor; C, cyclooxygenase-2) using random picking (black dotted lines); the HS-Pharm approach (black solid lines); and molecular docking with FlexX (red lines), Gold (green lines), and Surflex (blue lines). Identical targeted compound libraries and active site definitions were used for the comparison. ROC plots for HS-Pharm screens are incomplete since only compounds passing any of the four-feature pharmacophores (neuraminidase, P41 model; $\beta2$ receptor, P44 model; cyclooxygenase-2, P41 model) can be ranked by decreasing Fit values.

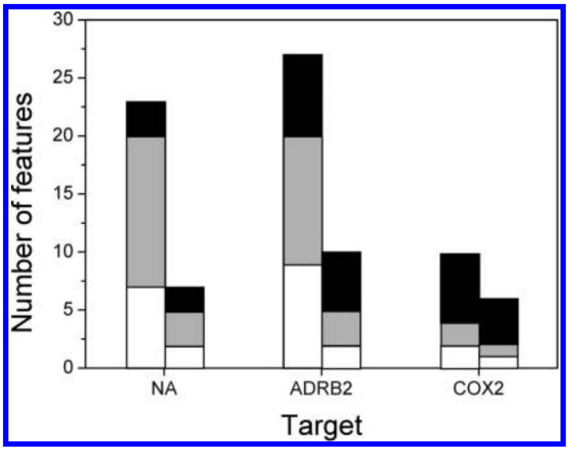

Figure 10. Number of pharmacophoric features defined from LUDI interaction maps, with (right columns) or without (left columns) machine-learning-based prioritization of ligand-interacting atoms. H-bond acceptor, H-bond donor, and hydrophobic features are represented by white, gray, and black bars, respectively.

Table 4. Number of Possible Pharmacophores P^a Depending on the Total Number of Features F and the Number of Features f Selected for the Query, with and without Restricting Feature Identification

	F			P	
protein	unrestricted[b]	restricted[c]	f	unrestricted	restricted
neuraminidase	23	7	3	1771	35
			4	8855	35
			5	33649	21
$\beta2$ receptor	27	10	3	2925	120
			4	17550	210
			5	80730	252
cyclooxygenase-2	10	6	3	120	20
			4	210	15
			5	252	6

[a] $P = F!/f!(F - f)!$. [b] Catalyst features derived from LUDI interaction maps. [c] Catalyst features derived from LUDI interaction maps, close to predicted ligand-interacting atoms according to the voting model on CFP2 cavity fingerprints.

practically screen a library of reasonable size (>10 000 ligands), even after 2-D filtering. It is thus of the utmost importance to reduce the number of interesting features and consequently the number of pharmacophore combinations without losing much information. Prediction of the most likely anchoring atoms with a machine learning algorithm and restriction of the pharmacophore definition only to those features close to these atoms considerably decreases the number of features by at least a factor 2 (Figure 10) and thus drastically simplifies the number of all corresponding pharmacophores (Table 4).

However, one has to be aware of the limitations of the HS-Pharm approach. There are still drawbacks to the proposed flowchart. The first one resides in the fact that the LUDI interaction maps consist only of three possible feature types (H-bond acceptor, H-bond donor, and hydrophobe) and omit charged and aromatic features, commonly used in ligand-based pharmacophore definitions. This simplification, while not affecting sensitivity, is likely to decrease the specificity of the receptor-based model since a H-bond acceptor can be mapped, for example, to what should be a negatively ionizable feature, or a hydrophobic group to what should be an aromatic feature, and thus lead to false positives in the hit list (Figure 11). Looking at the individual accessibility and probability of interaction (I_A) for atoms selected by our machine learning approach did not permit the derivation of general rules for modifying H-bond acceptor/donor and hydrophobe features into negative/positive ionizable and aromatic features. A manual editing of features is still possible, but it requires a significant knowledge of actives

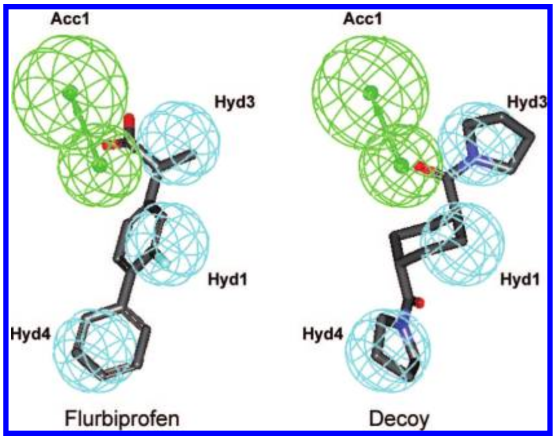

Figure 11. Mapping a true-positive COX-2 inhibitor (flurbiprofen, left panel) and a false-positive decoy (right panel) to one of the 15 four-feature HS-Pharm COX-2 inhibitor pharmacophores. Whereas both compounds have been selected for a good reason (four matches for each molecule), switching Acceptor 1 (Acc1) to a negatively ionized feature and Hydrophobe 4 (Hyd4) to an aromatic feature would penalize the false-positive (only two matches) without affecting the true-positive.

and their structure−activity relationships, and therefore it is not applicable for novel genomic targets.

A second drawback lies in the physicochemical properties of the binding site of interest. As far as polar and hydrophobic features are well-balanced (e.g., neuraminidase and β2 adrenergic receptor), the corresponding pharmacophores will capture the directionality of molecular interactions necessary for a ligand to achieve binding. For very hydrophobic binding sites (e.g., COX-2), this directionality is lost, since the ratio of hydrophobic-over-polar features is too high (Figure 10). The resulting pharmacophores (Figure 8) are then much less specific and slightly less sensitive than those derived from more polar binding sites (Figures 4 and 6). For strongly apolar binding sites, the herein proposed structure-based approach can thus be considered as a preliminary filtering step able to downsize a chemical library.

As expected for any structure-based method, HS-Pharm is sensitive to protein atomic coordinates and induced fit effects. This explains, in addition to the hydrophobic nature of the cavity, the poorer performance of the HS-Pharm screening in the case of the COX-2 target for which rotameric states of a few key anchoring residues are ligand-dependent. We have not currently compared HS-Pharm to other methods able to generate structure-based pharmacophores[11,17,23,25] but think that the above-reported conclusions still hold. The main advantage of the HS-Pharm approach does not reside in the pharmacophore perception and screening but in the significant reduction of the number of pharmacophoric features and possible pharmacophores thanks to the machine-learning prediction of hot spots. In our hands, restricting the number of features/pharmacophores increases the specificity of the pharmacophore searches by lowering the number of selected false positives. Therefore, applying the same preselection step to other methods will probably lead to very similar results. We chose to couple our knowledge-based ligand-anchoring prediction method to Discovery Studio for the possibility to script many operations in automated workflows (e.g., screening all *n*-feature pharmacophores, editing heat maps) that considerably facilitate exhaustive pharmacophore definitions and library screening.

CONCLUSIONS

We herewith present a structure-based approach to generate simple but efficient pharmacophore models by using a machine learning algorithm trained on known cavity fingerprints first to predict the most likely ligand-anchoring atoms and second to define receptor-based pharmacophores from probe interaction maps focusing on those atoms. The approach has been applied to three ligand-binding sites of pharmacological interest in order to distinguish known actives from chemically similar decoys. As already noticed for ligand-based pharmacophore searches,[70] four-feature pharmacophoric descriptions achieve the best compromise between sensitivity and specificity, whatever the cavity. Identification of features which are specifically matched by true positives with respect to positives led in all cases to models of significantly higher specificity at the cost of a slightly reduced sensitivity. Although we are lacking large-scale benchmarks, it appears that a difference higher than 15% in the above-cited matched features systematically improves the accuracy of pharmacophore searches. Of course, this additional postprocessing requires prior knowledge of known actives. However, even in the absence of known ligands, the general protocol presented in this study is sensitive and specific enough to prioritize virtual hits of interest. The main advantage of focusing pharmacophore features to previously identified anchoring atoms lies in the significant simplification of the resulting pharmacophoric description, which enables a systematic screening of all possible four-feature pharmacophores. Like in any pharmacophore search, some permissivity in the query is observed for unbalanced pharmacophores in which hydrophobic features are predominant. In the latter case, we advise the use of hot-spot-guided receptor-based pharmacophore searches as a preliminary filtering tool to downsize a compound library before *in silico* screening with another computational method.

ACKNOWLEDGMENT

We sincerely thank Hughes-Olivier Bertrand, Rémy Hoffmann, and Katalin Nadassy (Accelrys, Inc.) for numerous discussions and constructive comments on the manuscript. The European Union is acknowledged for granting C.B. with a Marie Curie Fellowship (MRTN-CT-2005-019561).

Supporting Information Available: Table S1: Definition of three cavity fingerprints (CFP1, CFP2, CFP3). Figure S1: Distribution of accessibility[31] for 122 070 interacting atoms (gray bars) and 501 689 noninteracting protein atoms (white bars) in 3500 protein−ligand sc-PDB entries.[27] Atoms are classified in three groups according to their accessibility (<5%, 5−30%, >30%). Figure S2: Chemical structures of eight neuraminidase inhibitors used to build a targeted library. Figure S3: Chemical structures of 20 β2 receptor ligands used to build a targeted library. Full agonists, partial agonists, inverse agonists, and neutral antagonists are indicated in red, orange, green, and blue, respectively. Please note that the functional effects of some of these compounds may vary with the signaling pathway. Figure S4: Chemical structures of 10 COX-2 inhibitors' used to build a targeted library. Chart S1: Literature references for neuraminidase inhibitors (refs 1−5), β2 receptor ligands (refs 6−22), and COX-2 inhibitors (ref 23). This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discovery* **2004**, *3*, 935–949.

(2) Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br. J. Pharmacol.* **2008**, *153*, S7–S28.

(3) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; 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.

(4) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **2004**, *57*, 225–242.

(5) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1999**, *42*, 5100–5109.

(6) 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.

(7) Carlson, H. A.; Masukawa, K. M.; Rubins, K.; Bushman, F. D.; Jorgensen, W. L.; Lins, R. D.; Briggs, J. M.; McCammon, J. A. Developing a dynamic pharmacophore model for HIV-1 integrase. *J. Med. Chem.* **2000**, *43*, 2100–2114.

(8) Fox, T.; Haaksma, E. E. Computer based screening of compound databases: 1. Preselection of benzamidine-based thrombin inhibitors. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 411–425.

(9) 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*, 2146–2157.

(10) Chen, J.; Lai, L. Pocket v.2: further developments on receptor-based pharmacophore modeling. *J. Chem. Inf. Model.* **2006**, *46*, 2684–2691.

(11) Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for Ligands and Proteins (FLAP): theory and application. *J. Chem. Inf. Model.* **2007**, *47*, 279–294.

(12) Wolber, G.; Seidel, T.; Bendix, F.; Langer, T. Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discovery Today* **2008**, *13*, 23–29.

(13) Bohm, H. J. The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 61–78.

(14) Phillips, G. N., Jr.; Fox, B. G.; Markley, J. L.; Volkman, B. F.; Bae, E.; Bitto, E.; Bingman, C. A.; Frederick, R. O.; McCoy, J. G.; Lytle, B. L.; Pierce, B. S.; Song, J.; Twigger, S. N. Structures of proteins of biomedical interest from the Center for Eukaryotic Structural Genomics. *J. Struct. Funct. Gen.* **2007**, *8*, 73–84.

(15) Kirchhoff, P. D.; Brown, R.; Kahn, S.; Waldman, M.; Venkatachalam, C. M. Application of structure-based focusing to the estrogen receptor. *J. Comput. Chem.* **2001**, *22*, 993–1003.

(16) *Discovery Studio*, version 2.0; Accelrys, Inc.: San Diego, CA.

(17) Schuller, A.; Fechner, U.; Renner, S.; Franke, L.; Weber, L.; Schneider, G. A pseudo-ligand approach to virtual screening. *Comb. Chem. High Throughput Screening* **2006**, *9*, 359–364.

(18) Kelly, M. D.; Mancera, R. L. A new method for estimating the importance of hydrogen-bonding groups in the binding site of a protein. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 401–414.

(19) Kelly, M. D.; Mancera, R. L. A new method for estimating the importance of hydrophobic groups in the binding site of a protein. *J. Med. Chem.* **2005**, *48*, 1069–1078.

(20) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.

(21) Verdonk, M. L.; Cole, J. C.; Taylor, R. SuperStar: a knowledge-based approach for identifying interaction sites in proteins. *J. Mol. Biol.* **1999**, *289*, 1093–1108.

(22) Bruno, I. J.; Cole, J. C.; Lommerse, J. P.; Rowland, R. S.; Taylor, R.; Verdonk, M. L. IsoStar: a library of information about nonbonded interactions. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 525–537.

(23) Ahlstrom, M. M.; Ridderstrom, M.; Luthman, K.; Zamora, I. Virtual screening and scaffold hopping based on GRID molecular interaction fields. *J. Chem. Inf. Model.* **2005**, *45*, 1313–1323.

(24) Spannhoff, A.; Heinke, R.; Bauer, I.; Trojer, P.; Metzger, E.; Gust, R.; Schule, R.; Brosch, G.; Sippl, W.; Jung, M. Target-based approach to inhibitors of histone arginine methyltransferases. *J. Med. Chem.* **2007**, *50*, 2319–2325.

(25) Ortuso, F.; Langer, T.; Alcaro, S. GBPM: GRID-based pharmacophore model: concept and application studies to protein-protein recognition. *Bioinformatics* **2006**, *22*, 1449–1455.

(26) Witten, I. H.; Frank, E. *Data mining. Practical machine learning tools and techniques*; Elsevier: Amsterdam, 2005.

(27) Kellenberger, E.; Muller, P.; Schalon, C.; Bret, G.; Foata, N.; Rognan, D. sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank. *J. Chem. Inf. Model.* **2006**, *46*, 717–727.

(28) 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*, 235–242.

(29) Marcou, G.; Rognan, D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.* **2007**, *47*, 195–207.

(30) *OEChem*, version 1.4.2; OpenEye Scientific software: Santa Fe, NM.

(31) *Molecular Surface Package*, version 3.0.3; Biohedron: Menlo Park, CA.

(32) *Pipeline Pilot*, version 6.1; SciTegic Inc.: San Diego, CA.

(33) Quinlan, J. R. *Programs for Machine Learning*; Morgan Kaufmann Publishers: San Francisco, CA, 1993.

(34) Breiman, L. Random forests. *Machine Learning* **2001**, *45*, 5–32.

(35) Freund, Y.; Shapire, R. E. In *Experiments with a new boosting algorithm*, Proceedings of the 30th International Conference on Machine Learning, San Francisco, 1996; Morgan Kaufmann: San Francisco, CA, 1996; pp 148−156.

(36) Rogers, D.; Brown, R. D.; Hahn, M. Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screen.* **2005**, *10*, 682–686.

(37) 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.

(38) *Filter*, version 2.0.1; OpenEye Scientific software: Santa Fe, NM.

(39) *JChem*, version 3.2.3; ChemAxon Kft.: Budapest, Hungary.

(40) *Corina*, version 3.4; Molecular Networks GmbH: Erlangen, Germany.

(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*, 1294–1303.

(42) *FlexX*, version 2.2; BiosolveIT GmBH: Sankt Augustin, Germany.

(43) *Surflex*, version 2.11; BioPharmics LLC: San Mateo, CA.

(44) *Gold*, version 3.2; The Cambridge Crystallographic Data Centre: Cambridge, U. K.

(45) *MOE*, version 2007.09; Chemical Computing Group: Montreal, Canada.

(46) Apache Tomcat. http://bioinfo-pharma.u-strasbg.fr/scPDB (accessed Apr 24, 2008).

(47) von Itzstein, M. The war against influenza: discovery and development of sialidase inhibitors. *Nat. Rev. Drug Discovery* **2007**, *6*, 967–974.

(48) Varghese, J. N.; Smith, P. W.; Sollis, S. L.; Blick, T. J.; Sahasrabudhe, A.; McKimm-Breschkin, J. L.; Colman, P. M. Drug design against a shifting target: a structural basis for resistance to inhibitors in a variant of influenza virus neuraminidase. *Structure* **1998**, *6*, 735–746.

(49) Taylor, M. R. Pharmacogenetics of the human beta-adrenergic receptors. *Pharmacogenomics J.* **2007**, *7*, 29–37.

(50) Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G.; Thian, F. S.; Kobilka, T. S.; Choi, H. J.; Kuhn, P.; Weis, W. I.; Kobilka, B. K.; Stevens, R. C. High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science* **2007**, *318*, 1258–1265.

(51) Swaminath, G.; Deupi, X.; Lee, T. W.; Zhu, W.; Thian, F. S.; Kobilka, T. S.; Kobilka, B. Probing the beta2 adrenoceptor binding site with catechol reveals differences in binding and activation by agonists and partial agonists. *J. Biol. Chem.* **2005**, *280*, 22165–22171.

(52) Salom, D.; Lodowski, D. T.; Stenkamp, R. E.; Le Trong, I.; Golczak, M.; Jastrzebska, B.; Harris, T.; Ballesteros, J. A.; Palczewski, K. Crystal structure of a photoactivated deprotonated intermediate of rhodopsin. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 16123–16128.

(53) 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*, 509–538.

(54) Xie, W. L.; Chipman, J. G.; Robertson, D. L.; Erikson, R. L.; Simmons, D. L. Expression of a mitogen-responsive gene encoding prostaglandin synthase is regulated by mRNA splicing. *Proc. Natl. Acad. Sci. U. S. A.* **1991**, *88*, 2692–2696.

(55) Masferrer, J. L.; Zweifel, B. S.; Manning, P. T.; Hauser, S. D.; Leahy, K. M.; Smith, W. G.; Isakson, P. C.; Seibert, K. Selective inhibition of inducible cyclooxygenase 2 in vivo is antiinflammatory and nonulcerogenic. *Proc. Natl. Acad. Sci. U. S. A.* **1994**, *91*, 3228–3232.

(56) Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.* **2006**, *49*, 534–553.

(57) Shepphird, J. K.; Clark, R. D. A marriage made in torsional space: using GALAHAD models to drive pharmacophore multiplet searches. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 763–771.

(58) Sperandio, O.; Andrieu, O.; Miteva, M. A.; Vo, M. Q.; Souaille, M.; Delfaud, F.; Villoutreix, B. O. MED-SuMoLig: a new ligand-based screening tool for efficient scaffold hopping. *J. Chem. Inf. Model.* **2007**, *47*, 1097–1110.

(59) Steindl, T.; Langer, T. Influenza virus neuraminidase inhibitors: generation and comparison of structure-based and common feature pharmacophore hypotheses and their application in virtual screening. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1849–1856.

(60) Zhang, Q.; Muegge, I. Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: ranking, voting, and consensus scoring. *J. Med. Chem.* **2006**, *49*, 1536–1548.

(61) Franke, L.; Byvatov, E.; Werz, O.; Steinhilber, D.; Schneider, P.; Schneider, G. Extraction and visualization of potential pharmacophore points using support vector machines: application to ligand-based virtual screening for COX-2 inhibitors. *J. Med. Chem.* **2005**, *48*, 6997–7004.

(62) Michaux, C.; de Leval, X.; Julemont, F.; Dogne, J. M.; Pirotte, B.; Durant, F. Structure-based pharmacophore of COX-2 selective inhibitors and identification of original lead compounds from 3D database searching method. *Eur. J. Med. Chem.* **2006**, *41*, 1446–1455.

(63) Palomer, A.; Cabre, F.; Pascual, J.; Campos, J.; Trujillo, M. A.; Entrena, A.; Gallo, M. A.; Garcia, L.; Mauleon, D.; Espinosa, A. Identification of novel cyclooxygenase-2 selective inhibitors using pharmacophore models. *J. Med. Chem.* **2002**, *45*, 1402–1411.

(64) Renner, S.; Schneider, G. Fuzzy pharmacophore models from molecular alignments for correlation-vector-based virtual screening. *J. Med. Chem.* **2004**, *47*, 4653–4664.

(65) Singh, S. K.; Saibaba, V.; Rao, K. S.; Reddy, P. G.; Daga, P. R.; Rajjak, S. A.; Misra, P.; Rao, Y. K. Synthesis and SAR/3D-QSAR studies on the COX-2 inhibitory activity of 1,5-diarylpyrazoles to validate the modified pharmacophore. *Eur. J. Med. Chem.* **2005**, *40*, 977–990.

(66) Rollinger, J. M.; Haupt, S.; Stuppner, H.; Langer, T. Combining ethnopharmacology and virtual screening for lead structure discovery: COX-inhibitors as application example. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 480–488.

(67) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.

(68) Lundstrom, K. Structural genomics and drug discovery. *J. Cell. Mol. Med.* **2007**, *11*, 224–238.

(69) McGovern, S. L.; Shoichet, B. K. Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. *J. Med. Chem.* **2003**, *46*, 2895–2907.

(70) 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.

CI800064Z