Development and Validation of a Novel Protein−Ligand Fingerprint To Mine Chemogenomic Space: Application to G Protein-Coupled Receptors and Their Ligands

Nathanael Weill and Didier Rognan*

Structural Chemogenomics Group, Laboratory of Therapeutic Inovation, UMR 7200 CNRS-UdS (Université de Strasbourg), 74 route du Rhin, B.P.24, F-67400 Illkirch, France

Received December 11, 2008

The present study introduces a novel low-dimensionality fingerprint encoding both ligand and target properties which is suitable to mine protein−ligand chemogenomic space. Whereas ligand properties have been represented by standard descriptors, protein cavities are encoded by a fixed length bit string describing pharmacophoric properties of a definite number of binding site residues. In order to simplify the cavity fingerprint, the concept was applied here to a unique family of targets (G protein-coupled receptors) with a homogeneous cavity description. Particular attention was given to set up data sets of really diverse protein−ligand pairs covering as exhaustively as possible both ligand and target spaces. Several machine learning classification algorithms were trained on two sets of roughly 200000 receptor−ligand fingerprints with a different definition of inactive decoys. Cross-validated models show excellent precision (>0.9) in distinguishing true from false pairs with a particular preference for support vector machine classifiers. When applied to two external test sets of GPCR ligands, the most predictive models were not those performing the best in the previous cross-validation. The ability to recover true GPCR ligands (ligand prediction mode) or true GPCRs (receptor prediction mode) depends on multiple parameters: the molecular complexity of the ligands, the chemical space from which ligand decoys are selected to generate false protein−ligand pairs, and the target space under consideration. In most cases, predicting ligands is easier than predicting receptors. Although receptor profiling is possible, it probably requires a more detailed description of the ligand-binding site. Noteworthy, protein−ligand fingerprints outperform the corresponding ligand fingerprints in mining the GPCR-ligand space. Since they can be applied to a much larger number of receptors than ligand-based fingerprints, protein−ligand fingerprints represent a novel and promising way to directly screen protein−ligand pairs in chemogenomic applications.

INTRODUCTION

Chemogenomics is a novel research area aimed at identifying all possible ligands of all possible targets.[1−3] Since the corresponding target-ligand interaction matrix cannot be fully filled at the experimental level despite noticeable efforts,[4] computational chemistry and computational biology methods are supposed to clearly enhance the predictive power of *in silico* chemogenomic applications.[5] Predictive chemogenomic methods generally rely on either comparing biologically annotated ligands properties[6−11] or ligand-annotated receptors (binding sites) properties.[12−14] Finding novel ligands for a given target or a novel target for a given ligand is simply inferred from the basic principle that *similar receptors bind to similar ligands*.[15] Mining approaches in which both protein and ligand spaces are addressed with the same descriptor are still rare.[16−20] An implication for this observation is that such true chemogenomic approaches require a unique framework for comparing ligand-binding sites of various sizes. Up to now, most applications[18−20] have focused on G protein-coupled receptors (GPCRs) for various reasons: (i) GPCRs still represent the main family of targets for drug discovery;[21] (ii) a large array of experimental data on known GPCR-ligand pairs is available;[22,23] (iii) GPCR ligands readily cross-react with unrelated GPCRs;[24] (iv) GPCR cavities can be simplified to a fixed array of binding site residues[25,26] thus simplifying the definition of a generic cavity descriptor while rendering the data analysis more difficult because of the limited diversity of GPCR space covering only about 400 human nonolfactive receptors;[27]and (v) such approaches do not require three-dimensional (3-D) protein structure information, which are still exceptionally difficult to gather for this family of membrane proteins.[28]

Two previous studies on fingerprinting GPCR-ligand pairs in chemogenomic applications have been described. In a pioneering work, Bock et al.[20] used rather standard 2-D topological and atomic descriptors for ligands, physicochemical properties of amino acid sequences for receptors, and concatenate feature vectors for both the receptor and the ligand in a single fingerprint. A Support Vector Machine (SVM) model was trained on 5319 receptor−ligand pairs from the PDSP K_i database[22] to predict the K_i of any ligand to any GPCR and used to propose novel ligands for orphan GPCRs. Unfortunately, none of these predictions have been validated up to now. Recently, Jacob et al.[19] proposed a similar approach on 4051 pairs from the GLIDA database[23] with the noticeable exception that the tensor product between vectors describing ligands and proteins were used to better delineate correlations between ligand and target features. A

* Corresponding author phone: +33-3-90244235; fax: +33-3-90244310; e-mail: didier.rognan@pharma.u-strasbg.fr.

10.1021/ci800447g CCC: $40.75
Published on Web 03/20/2009

SVM classifier was used to train and predict out-of-sample pairs, but no convincing external test cases could be provided.

Both studies present however the remarkable advantage to unambiguously demonstrate that GPCR-ligand pairs can be encoded by a single vector and that machine learning classifiers can recover true receptor−ligand pairs. However, the general applicability of both models is unknown since the number of pairs on which they have been trained on is below our current knowledge on GPCR-ligand interactions. Moreover, the influence of key parameters (knowledge of the ligand binding site, selection of inactive decoys, machine learning algorithm, external test set validation) were not exhaustively addressed in these two seminal studies.

We herewith present a novel protein−ligand fingerprint (PLFP) describing pharmacophoric properties of ligands and their respective transmembrane binding cavities and its application to mine GPCR chemogenomic space. Various machine learning classifiers have been trained, on the most exhaustive and diverse set of receptor−ligand pairs gathered up to date, to predict the binary association of any individual ligand to any individual receptor. The influence of several parameters on the true predictive accuracy of these models was ascertained by using two external test sets of 60 GPCR ligands. Practical guidelines are given to favor the experimental validation of *in silico* predictions to find either novel ligands for a particular target or novel targets for a particular ligand.

METHODS

Setting up Data Sets of GPCR-Ligand Pairs. GPCR ligands were retrieved from the 2007.2 release of the *MDL Drug Data Report*.[29] First, a list of 96 activity class numbers corresponding to unambiguous GPCR targets was manually collected and annotated by SwissProt entry name (Supporting Information Table S1). Corresponding molecules were collected in SD file format, ionized at physiological pH with Filter2[30] (parameters in Supporting Information Table S2) and standardized for structure homogeneity with Standardizer v5.0.0[31] (parameters in Supporting Information Table S3). Peptides and duplicates were last removed with an in-house Pipeline Pilot script.[32] A set of 21050 unique ligands was selected and biologically annotated at the receptor level, describing a total of 160 human GPCR entries. In the event that no receptor subtype was explicitly mentioned in the MDDR database (e.g., galanin antagonists), all subtypes of the given receptor were assumed to bind to the corresponding ligand (see complete annotation in Supporting Information Table S1). A total of 32118 GPCR-ligand pairs could be identified. When a receptor subtype was annotated by more than 100 unique ligands, a MaxiMin distance algorithm[33] was applied to all ligands of the same receptor using pairwise Tanimoto similarity coefficients from MACCS structural keys generated in MOE (version 2007.09),[34] and the 100 most dissimilar ligands were finally selected for each receptor. The final number of dissimilar GPCR-ligand pairs was 8250. For each of the 160 GPCR entries, a set of presumed inactive decoys was generated from our in-house data set of 1.8 million commercially available druglike compounds.[35] Decoys were selected to span the same molecular weight range than all true actives for a given receptor and to be distant enough (Tanimoto coefficient below 0.75 on MACCS structural keys) from any of these actives. The MaxiMin clustering algorithm was used to select the most chemically dissimilar decoys in order to ensure, for each receptor entry, a constant balance between true pairs or actives (5%) and false pairs or inactives (95%). A first data set (**Data set 1**) of 168536 GPCR-ligand pairs (8250 actives and 160286 inactives) was thus finally designed.

A second data set (**Data set 2**) of pairs was generated considering all 32118 GPCR-ligand pairs but using a different set of decoys than that previously described for Data set 1. In Data set 2, decoys are selected from the MDDR GPCR ligand space. Decoys for receptor R_i are chosen if two conditions are verified: (i) the decoy is annotated as a ligand of receptor R_j but not of receptor R_i and (ii) R_j is distant enough from R_i (minimum Euclidean distance of 7.0) when the corresponding transmembrane binding cavities are compared using CavFP descriptors (see the "GPCR cavity descriptor" section below). Data set 2 was finally composed of 234137 GPCR-ligand pairs (32118 active and 202019 inactive pairs).

Validation Set of GPCR Ligands. Two sets of ligands targeting the corticotropin-releasing factor 1 receptor (CRFR1) and the neurokinin-1 receptor (NK1R) were extracted from the MDDR. For each activity class, a maximal diversity selection of 50 compounds was done using an in-house Pipeline Pilot script using MACCS public keys as descriptor. Property ranges (molecular weight, H-bond donor and acceptor counts) were computed for both sets of actives and used to select all bioactive ligands from the DUD database[36] fulfilling the requested property ranges. A set of 950 diverse DUD ligands was then selected as decoys.

External Test Sets of GPCR Ligands. Two external test sets of known GPCR ligands were used to validate classification models. The first set (Test set 1, Supporting Information Table S4) comprises 35 endogenous nonpeptide GPCR ligands (targeting 88 GPCR entries) not described in the MDDR data set. The second set (Test set 2, Supporting Information Table S5) was composed of 25 recently described synthetic "druglike" ligands, manually selected from the literature to bind to a diverse set of 28 GPCR targets.[37−60]

Fingerprinting GPCR Ligand-Binding Cavities (CavFP Fingerprint). For each of the 363 nonolfactive human GPCR targets considered in this work, a discontinuous sequence of 30 cavity-lining residues was retrieved, as previously described.[26] Each cavity was represented by a fixed-length vector of 240 bits describing every residue by 8 bits according to its pharmacophoric properties (H-bond donor, H-bond acceptor, charge, aromatic/aliphatic character, size; Supporting Information Table S6). To measure the pairwise distance between all GPCR entries, a global distance matrix was derived from Euclidean distances measured from CavFP fingerprints. The matrix was then converted into a phylogenetic tree using the UPGMA algorithm[61] as previously described.[26] A consensus final tree was built up from 1000 bootstraps using the CONSENSE program from the PHYLIP v3.2 suite.[62]

Fingerprinting GPCR Ligands. Since the ligand fingerprint needs to be later incorporated into a PLFP, the choice of possible descriptors was limited here to fixed-length vectors which size should be comparable to that encoding the cavity. Three different descriptors were thus considered to describe the ligand. First, the 166-bit public MACCS structural keys were computed in MOE[34] from 2-D SD structures of all 21050 unique GPCR ligands stored in the

database. Second, a modified version of the recently described SHED descriptor (SHannon Entropy Descriptor)[63] was implemented. Two additional pharmacophoric properties (positive charge, negative charge) were added to the original four properties (H-bond acceptor, H-bond donor, aromatic, apolar) resulting in 21 possible atom-centered feature pairs. The feature mapping to atom type was reconsidered to better account for charged and aromatic atoms (see complete mapping table in Supporting Information Table S7). Feature mapping was done by using the OEChem v1.4.2 library[30] and TRIPOS mol2 atom names.[64] The entropy SHED value of each of these 21 pairs was computed as originally described[63] and recorded into a vector of 21 reals.

Last, the third descriptor was directly derived from the above-described SHED descriptor by storing the full distribution (DistFP) of path lengths between all 21 feature pairs. Only the shortest path between two features was considered up to a maximum length of 18 bonds. The DistFP descriptor is a vector of 376 (21 * 18) reals.

Concatenation of GPCR and Ligand Fingerprints into a PLFP. The GPCR-ligand fingerprint was defined according to the approach proposed by Bock et al.[20] by concatenating the CavFP cavity descriptor with each of the three previously described ligand descriptor (MACCS, SHED, DistFP). A final bit was added to describe whether the corresponding GPCR-ligand binary association is registered in our MDDR-derived GPCR-ligand database or not.

Machine Learning Methods. Random Forest (RF) and naïve Bayesian (NB) classification models were generated using the Java machine learning workbench WEKA v3.5.5.[65] Default settings of the Breiman implementation[66] of RF in Weka was used, except for the maximal number of trees which was set to 50. SVM classification models were computed with libSVM v2.84.[67] All SVM models were built using a RBF kernel. Cost (C) and gamma (γ) parameters were optimized for each model by testing 32 possible combinations ($C = 10^{[0:3]}$, $\gamma = 2^{[-6:2]}$). Whatever the variation (machine learning method, fingerprint, decoy set), the SVM model finally selected was the one that leads to the best balanced accuracy Ba

$$Ba = \frac{TP/P + TN/N}{2}$$

where TP = number of true positives, TN = number of true negatives, P = number of positives, and N = number of negatives.

The criteria used to compare classification models are recall, precision, and F-measure, which are defined as follows

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F = \frac{2(Precision \times Recall)}{Precision + Recall}$$

where FN = number of false negatives, and FP = number of false positives.

RESULTS AND DISCUSSION

The aim of the current study is to evaluate the relevance of machine learning classification models for predicting binary association between a GPCR and a ligand from simple 1-D protein−ligand fingerprints. If predictions have to be tested experimentally, one would restrain the validation to a single receptor (ligand screening) or a single ligand (receptor profiling). *In silico* predictions were therefore performed in two directions: (i) recover/find ligands for a given receptor (from here on stated as "ligand mode") and (ii) recover/find targets for a given ligand (from here on stated as "receptor mode"). In order to carefully validate both the fingerprint and the classifying method, we will discuss throughout this section the influence of a single modification on various models. We should here recall that the aim of the current study is not to predict either binding modes or binding affinity but simply to anticipate whether or not a ligand may bind to the canonical transmembrane domain[26] of a GPCR.

Input Data and Fingerprints. Particular attention has been given in gathering a really representative and diverse set of GPCR annotated druglike ligands. The MDDR data set[29] was chosen as the starting data warehouse since it contains one of the largest collections of biologically annotated druglike compounds, which has very often been used as a source for developing/comparing GPCR ligand-based *in silico* screening methods.[68,69] Although the annotation at a molecular target level is not always straightforward with respect to other commercially available biologically annotated compound collections,[5] we do not believe that the choice of this particular data set leads to a significant bias in the results reported herein. A set of 21050 druglike compounds with a precise GPCR annotation, resulting in 32118 GPCR-ligand binary pairs, could be retrieved.

Importantly, peptides and peptidomimetics were discarded from the final selection since they are well-known to bind to the extracellular loops[70] and not the transmembrane cavity our CavFP fingerprint is focused on. Analogously, ligands binding to the extracellular domain of Class B and class C GPCRs[21] were also discarded manually. The GPCR biological space covered by our data set comprises 160 human receptors (Figure 1A), out of which 70 entries are described by at least 100 unique ligands. Nineteen out of 22 GPCR subfamilies, that we previously defined by comparing all nonolfactive human GPCRs with a structural chemogenomic approach,[26] are described in our data set (Figure 1B). Only three subfamilies of orphan receptors (Adhesion, MAS, SRBs)[26] are not addressed herein. As expected, the subfamily of biogenic amine receptor ligands is the most represented (39% of ligands) since it just reflects drug discovery trends over the last decades.[71] However, limiting the maximum number of ligands for each receptor to an upper value of 100 decreases the risk of biasing the ligand data set for a particular GPCR space. All druggable GPCR subfamilies (e.g., chemokine, purines, peptides) are currently addressed in the present study, even that for which the first nonpeptide ligands have been reported quite recently.[72]

The generic cavity descriptor used to describe GPCR transmembrane binding sites slightly varies from the one recently used in our chemogenomic analysis of GPCR cavities.[26] Instead of using concatenated binding site sequences and sequence identity as a pairwise distance, we

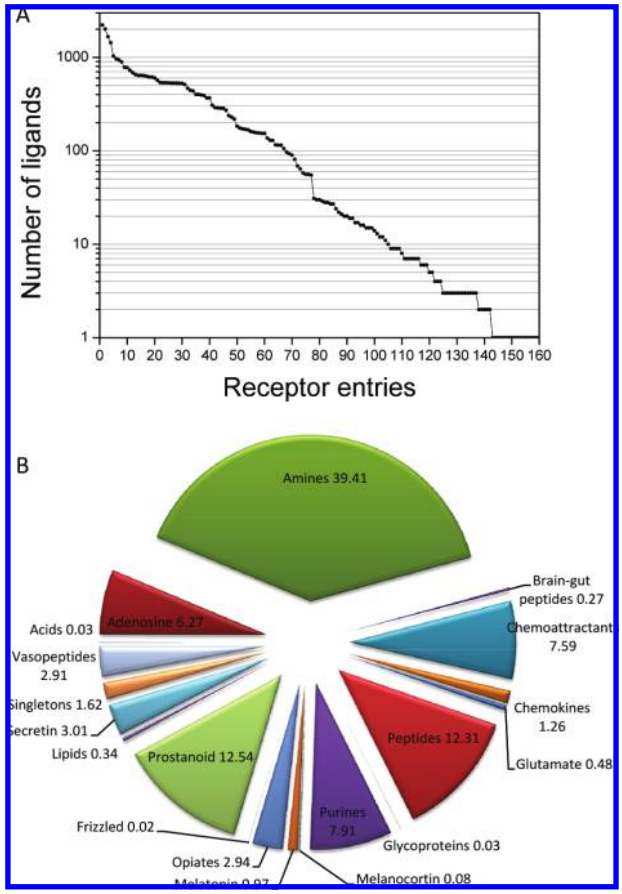

Figure 1. Coverage of ligand and receptor space by the GPCR-ligand data set. **A)** Number of unique ligands annotated for 160 human GPCRs. **B)** Ligand distribution (in percentage) for 19 out of 22 GPCR clusters described by Surgand et al.[26]

here encode in a bit string the pharmacophoric properties of each of the 30 residues[26] selected for defining a consensus ligand-binding site in the TM cavity. The relevance of the CavFP descriptor is outlined by its capacity to lead to a cavity-biased phylogenetic tree (Figure 2) in line with the full sequence-derived tree.[73] Only the family of 'Peptides GPCRs' (receptors for peptide endogenous ligands) is divided in two unrelated branches. For the remaining 21 subfamilies previously identified by sequence analysis,[26] receptors for specific chemotypes (e.g., purines, chemokines, lipids) are unambiguously grouped together (Figure 2).

Having validated the 'cavity block' of our PLFP, we looked next at the relevance of the 'ligand block'. SHED and MACCS descriptors have both been validated as information-rich molecular descriptors in previous reports.[74,75] Since we significantly modified the original SHED descriptor,[63] we challenged our SHED implementation as well as the derived DistFP descriptor (see the 'Methods' section) for discriminating true actives from chemically similar decoy ligands in a simple ligand-based classification model. Ligands of two unrelated GPCRs (CRFR1, NK1R) were retrieved from the MDDR data set. Decoys were chosen from bioactive compounds (mostly enzyme inhibitors) of the DUD database[36] presenting the same property range (molecular weight, H-bond donor and acceptor counts) than true CRFR1 and NK1R ligands. Three machine learning algorithms (NB, J48, and RF)) were applied to classify DUD ligands using three fingerprints (SHED, DistFP, MACCS). The accuracy of the classification (Figure 3) was assessed by computing the area under the ROC curve for the 9 possible combinations, by repeating 10 times a 10-fold cross-validation on each data set. The three ligand descriptors are indeed suitable to distinguish true actives from chemically similar decoys in both activity classes with ROC values between 0.7 and 0.96. In addition to the well investigated MACCS keys, the distFP descriptor, despite its relatively small size (378 reals), is remarkably accurate, whatever the machine learning algorithm, although extensive validation on much more activity classes would be required to draw more general conclusions. A clear trend is however that RF seems the best classification method, irrespective of the ligand descriptor used in both data sets (Figure 3). Since the J48 decision tree method was clearly inferior to both RF and NB models, it was discarded for the later analysis of protein−ligand fingerprints.

Classification and Prediction Models from Data Set 1. *Cross-validation Models.* Data set 1 is composed of 168536 GPCR-ligand pairs (8250 actives and 160286 inactives) in which decoys have been chosen within the chemical space of commercially available druglike compounds that do not intersect with the MDDR GPCR ligands. We first evaluated the performance of a global model versus local models by repeating 10 times a 10-fold cross-validation classification. Whereas the global model considers all data, 19 local models were done for each of the 19 GPCR subfamilies/clusters[26] for which GPCR-ligand pairs are available. The probability of occurrence of a protein−ligand pair is then given by the local model to which the corresponding receptor belongs to. Probabilities higher than 0.5 from all local models are then merged and ranked by decreasing values.

Three algorithms (SVM, RF, NB) were used in combination with three ligand descriptor blocks (SHED, DistFP, MACCS) in the local models. Only the SVM and NB classifications could be applied to the global model since the RF implementation in Weka could not load all data in memory. For all SVM models, the three descriptors and the two machine learning algorithms performed relatively well in terms of recall and F-measure (values between 0.6 and 0.85; Figure 4) and were excellent with respect to the precision in predicting true active GPCR-ligand pairs (values between 0.9 and 0.95). The DistFP/SVM combination on local models exhibits the peak performance and the lowest variability in the observed quality criteria (recall, precison, F-measure). Despite acceptable recall values, NB used as a classifier is significantly less interesting than SVM because of its much lower precision in both the global and local models (Figure 4). We anticipate that the 'cavity block' of the PLFPs is not variable enough for NB models, notably when applied to a single protein family as it is the case in the current study. Therefore, NB models were not used to predict protein−ligand pairs for the two external test sets.

Predicting GPCR-Ligand Pairs. Whether there is a correlation between the performance of a classification model in the latter cross-validation experiment and its true predictive power was addressed next by trying to predict either the true receptor(s) of GPCR ligands (receptor prediction mode) or the true ligand(s) of GPCRs (ligand prediction mode). For that purpose, ligands from two external test sets (Set 1: 35 endogenous nonpeptide ligands; Set 2; 25 synthetic ligands) absent from the original training set were used. Test set 1 is an exhaustive collection of all currently known nonpeptidic

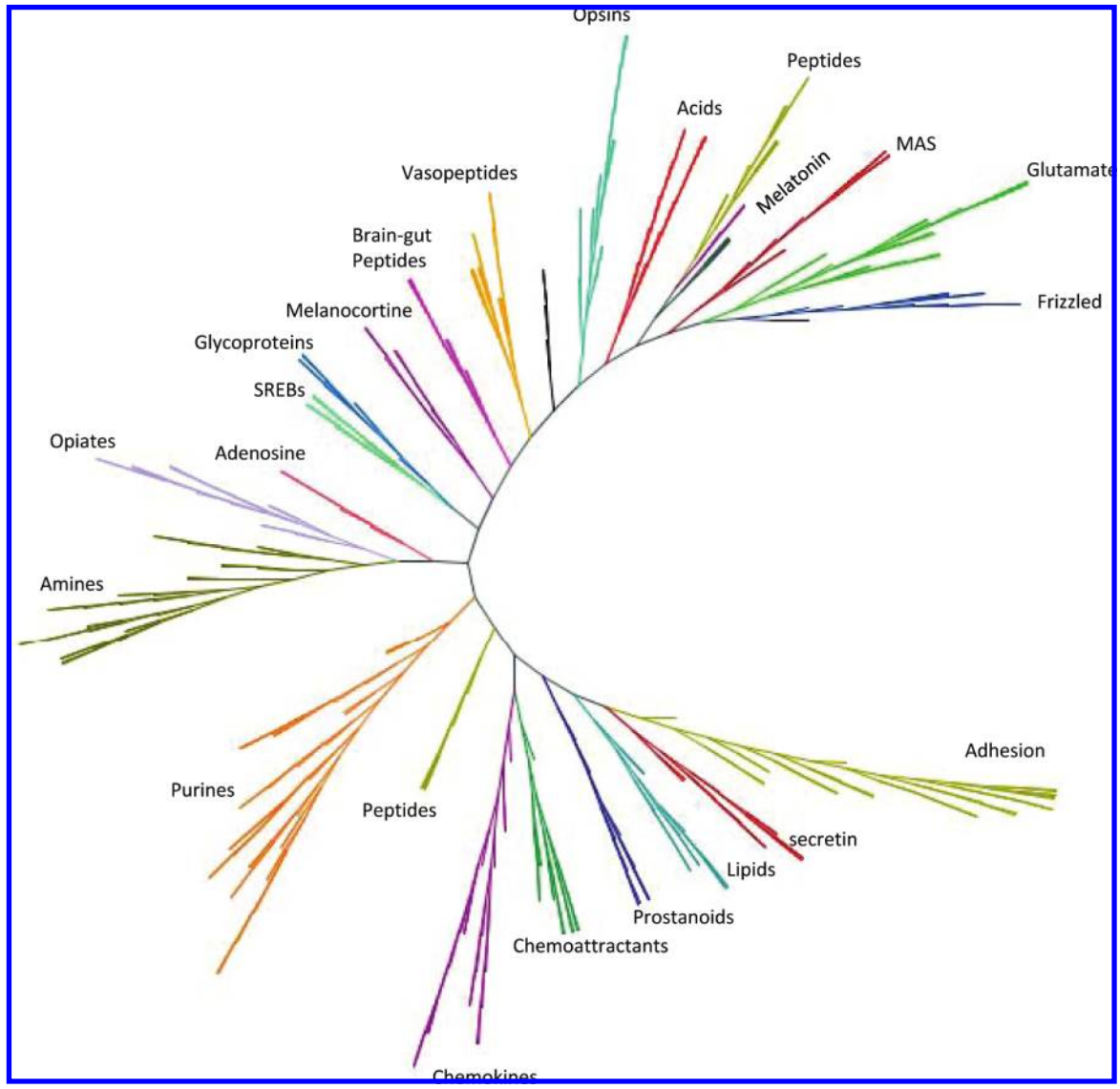

Figure 2. Phylogenetic tree of 363 human nonolfactive GPCRs, derived from an Euclidian distance matrix computed on GPCR cavity (CavFP) fingerprints. The tree has been computed and rendered with HyperTree.[84]

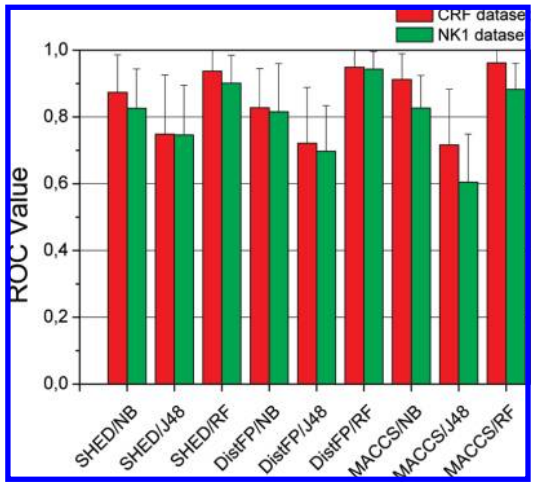

Figure 3. Performance (ROC score) of three machine learning models (Naïve Bayesian, NB; Decision trees, J48; Random Forest, RF) in classifying true actives and DUD decoys[36] of two data sets (cyclooxygenase-2 inhibitors, cox2, p38 MAPK kinase inhibitors, p38) from three molecular descriptors (SHED, DistFP, MACCS; see Computational Methods).

GPCR endogenous ligands and is therefore of interest to check whether our PLFPs are suitable to deorphanize a GPCR of interest. Test set 2 is a handmade collection of recently described synthetic ligands chosen for their structural novelty and target diversity (see Supporting Information Tables S4 and S5 for structure and target annotation). Each test set was supplemented by a set of decoys as follows (Figure 5). In ligand prediction mode, every true ligand was supplemented by its own collection of chemically similar commercially available decoys (1 active for 528 decoys in test set 1; 1 active for 549 decoys in test set 2) thus leading to 35 different external subsets (endogenous ligands) and 25 different external subsets (synthetic ligands). The receptor description within each subset is kept constant in the PLFP. In receptor prediction mode, decoys were simply any of the 363 GPCRs not associated with the true ligand in our database, and the ligand descriptor block then remained constant for each subset (Figure 5). In all cases, the corresponding ligand and GPCR cavity fingerprints were then concatenated and classified with the global model and the local models corresponding to the GPCR targeted by each of the actives (ligand prediction mode) or the local model

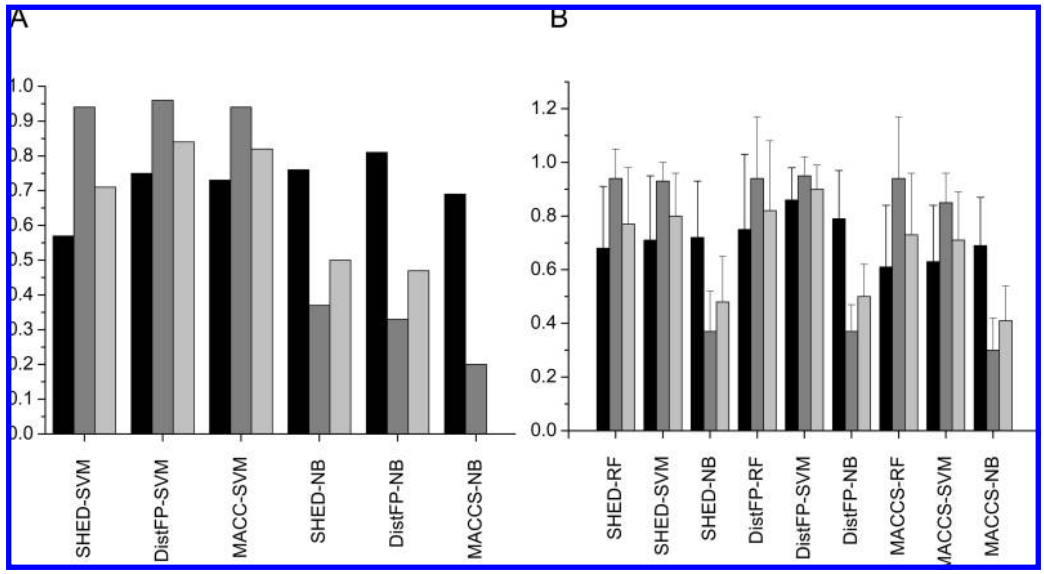

Figure 4. Performance of a global (panel A) and 19 local models (panel B) in classifying and predicting GPCR-ligand pairs from Data set 1. The mean value and standard deviation is given for all local models outputting a probability of binary association higher than 0.5. Criteria used to judge the performance are recall (dark gray bars), precision (gray bars), and F-measure (light gray bars).

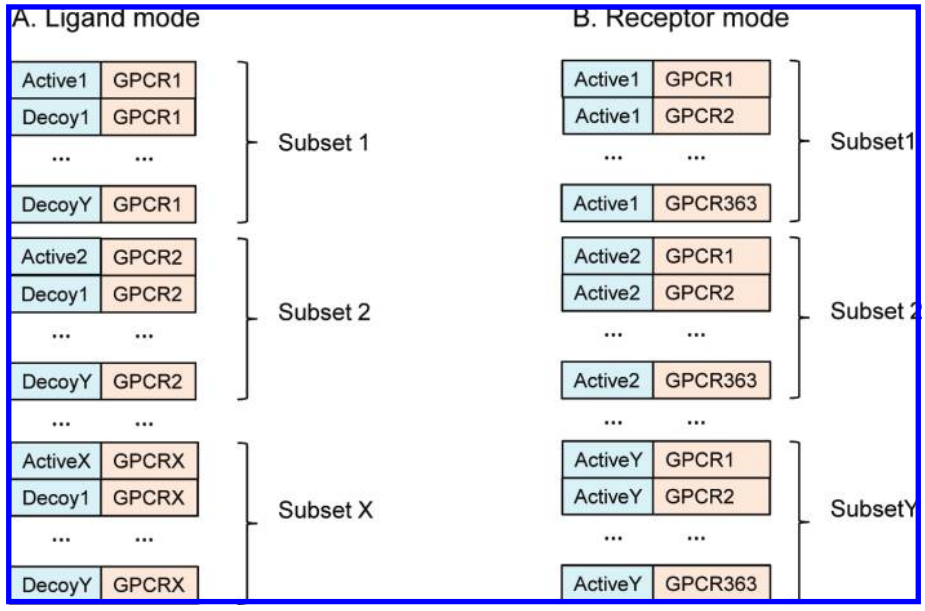

Figure 5. Construction of decoys sets in receptor prediction mode and ligand prediction mode. **A)** In ligand prediction mode, each true ligand is supplemented by druglike decoys (X=87, Y=527 for test set 1; X=28, Y=549 for test set 2) in a receptor-specific subset. Predictions are made from PLFPs where the ligand part (SHED, DistFP, MACCS; displayed as a light blue block) in each subset (Active 1 to Active X) is variable and the receptor part (displayed as an orange block) is constant. **B)** In receptor prediction mode, each true receptor is supplemented by GPCR decoys (Y = 35 for test set1, Y=25 for test set 2) in a ligand-specific subset. Predictions are made from PLFPs where the ligand part (SHED, DistFP, MACCS; displayed as a light blue block) in each subset (Active 1 to Active Y) is constant and the receptor part (displayed as an orange block) is variable.

corresponding to each of the 363 GPCRs described in the PLFP (receptor prediction mode).

The external validation was analyzed in order to answer two basic questions: (i) what is the size of the hit list (a hit being either a ligand or a target according to the prediction mode) and (ii) what is the rank of the true ligand/target in the hit list. Ideally, the virtual screen should lead to the smallest possible hit list and the best possible rank of the true hit.

In ligand prediction mode (Figure 6A, B) the best performance for predicting the 35 endogenous ligand of 88 different GPCRs was obtained with local models using the MACCS/RF combination. This model is very accurate in terms of recall (0.82) and precision (very short hit list and very low rank of the true endogenous ligand). The DistFP/RF combination is also appropriate for this test set. Other combinations are less interesting mainly because of a low specificity (large hit list but poor rank of the true ligand) although the recall is in all cases quite acceptable (0.74−0.85; Figure 6B). The global model is not appropriate whatever the machine learning method and the ligand descriptor mostly because of its low precision (Figure 6A).

Using the best classification model (MACCS/RF with local models), the endogenous ligand is ranked first in the hit list in 26 out of 35 hit lists (all results are summarized in Supporting Information Table S8). This overall excellent

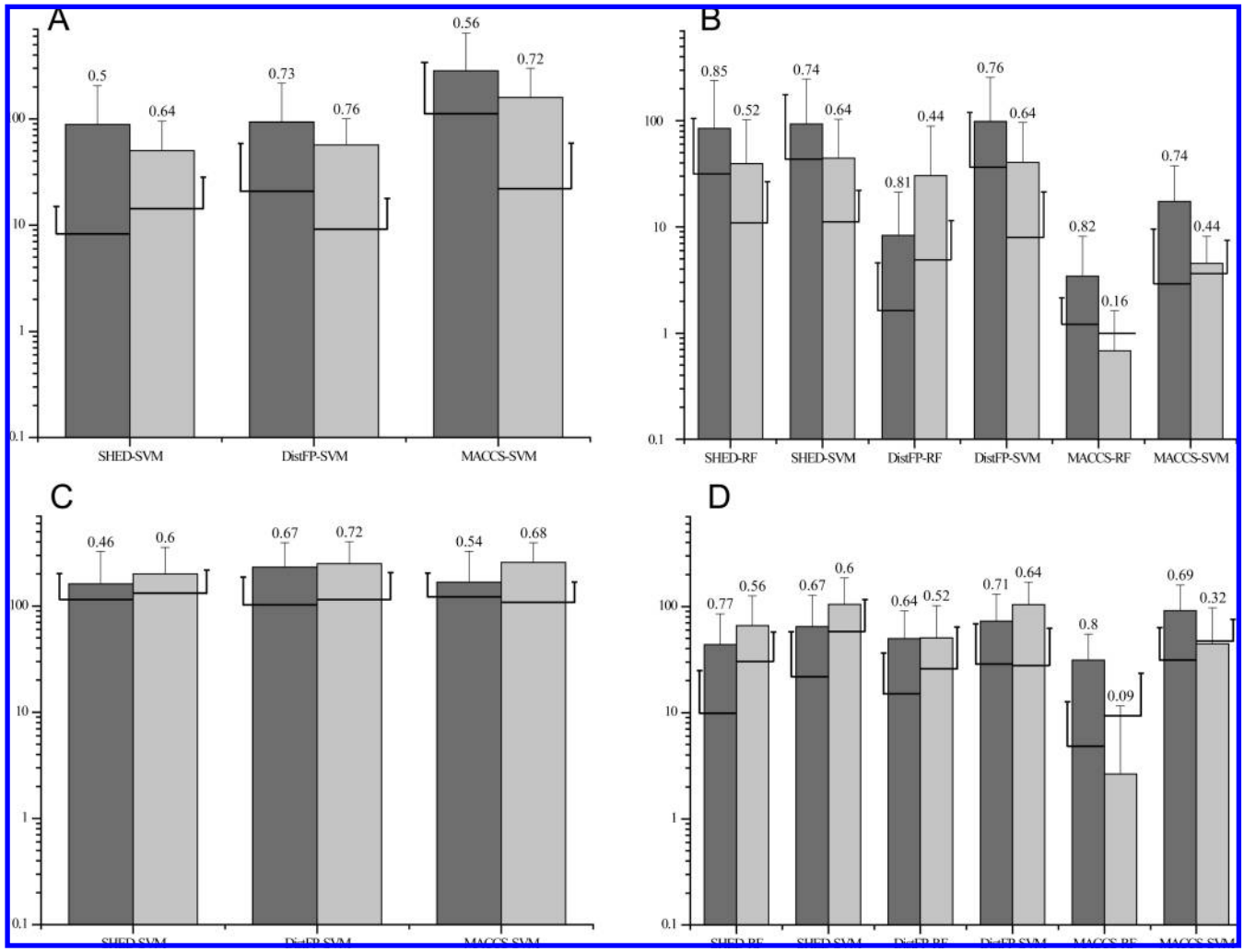

Figure 6. Predicting possible protein−ligand pairs for two external test sets (test set 1 of 35 endogenous GPCR ligands, dark gray bars; test set 2 of 25 synthetic GPCR ligands, light gray bars) with the global model (panels A and C) and the 19 local models (panels B and D) trained on Data set 1 (see Methods). The size of the ligand (panels A, B) or the receptor hit list (panel C, D) is shown by a bar, and the rank of the true ligand/receptor in the hit list is displayed by a horizontal bar-crossing line. Vertical lines starting from the bar and the bar-crossing line indicate the standard deviation of the hit list size and of the true hit rank, respectively. The number on the top of the bar is the recall of the prediction on the entire test set.

performance demonstrates that PLFPs are indeed well suited for predicting binary associations between a variable ligand and a fixed receptor. No predictions could be made in 7 cases (5-oxo ETE for OXOER1, kynurenic acid for GPR35, beta-alanine for MRGRD, leukotriene C4 for CLTR2, leukotriene D4 for CLTR1, propionic acid for FFAR2 and FFAR3, thromboxane B2 for GPR44). These cases correspond to situations where either very few data are available in the training set (e.g., ligands for FFAR1, FFAR2, OXOER1, GPR35, and MRGRD are absent or quite rare) or the synthetic MDDR ligands differ significantly from the endogenous compound (e.g., ligands for CLTR1, CLTR2, and GPR44). The test set 1 is far from being an easy external test set since it contains only full agonists, whereas synthetic ligands on which the models have been trained on are mostly neutral antagonists and inverse agonists. However, the results obtained with the best model are quite encouraging. It can be stated at this point that the best predictive model for test set 1 (local models with MACCS/RF) is not that performing the best in the previous cross-validation experiment on the training set (Figure 4). Analogously, the performance of the various models fluctuates dramatically when applied to the second external test set of 25 synthetic ligands. In the later case, recall values are significantly lower (0.16−0.64). The local models using the MACCS/RF combination are still interesting since they are quite specific and accurate but suffer from a low recall (0.16). In other words, this classification model rarely leads to predictions, but the few ones which are outputted are of very high quality. The best compromise for predicting true synthetic ligands is still achieved by local models but using DistFP as ligand descriptor and SVM as classification algorithm (Figure 6B). For 16 out of the 25 synthetic ligands in test set 2, the true ligand when seeded with a set of 549 chemically similar decoys is recovered in the hit list for its cognate GPCR entry. For many compounds (e.g., PRED1, PRED2, PRED10; Figure 7), the true ligand is ranked first within a short hit list (4% of the full list). In some cases, the model returned an empty list (e.g., PRED8; Figure 7) because the query ligand is too dissimilar to any known ligand of that receptor. The size of the hit list depends on the promiscuity of the corresponding chemotype for various GPCRs. If the chemotype is relatively selective for a particular GPCR subspace, the hit list is short and the true ligand ranked high (e.g., PRED1). However, if the chemotype

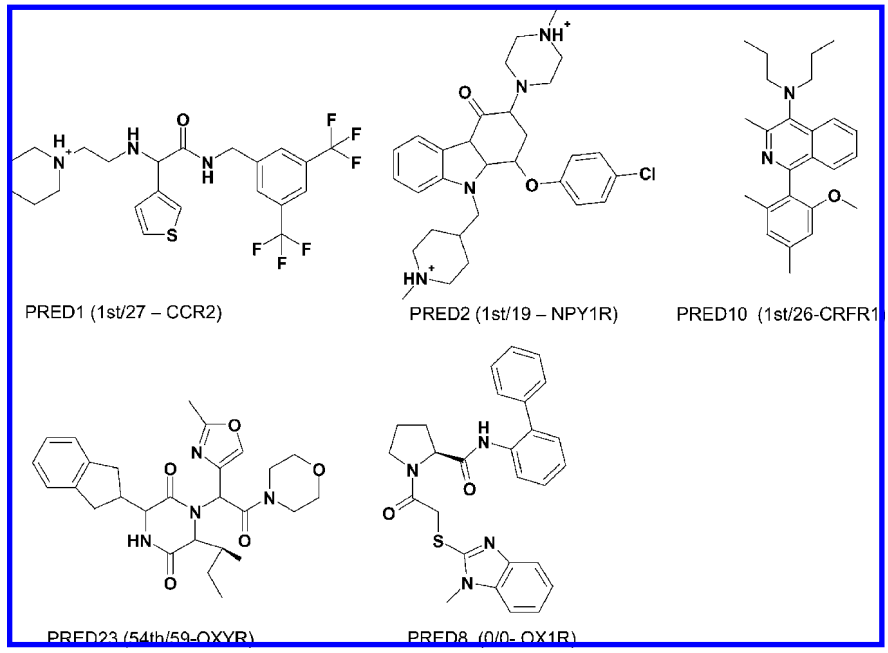

Figure 7. Examples of ligand predictions for various synthetic ligands of the external test set 2, using the DistFP-SVM model trained on Data set 1. Numbers in brackets indicate the rank of the corresponding true ligand and the size of the hit list for a given receptor.

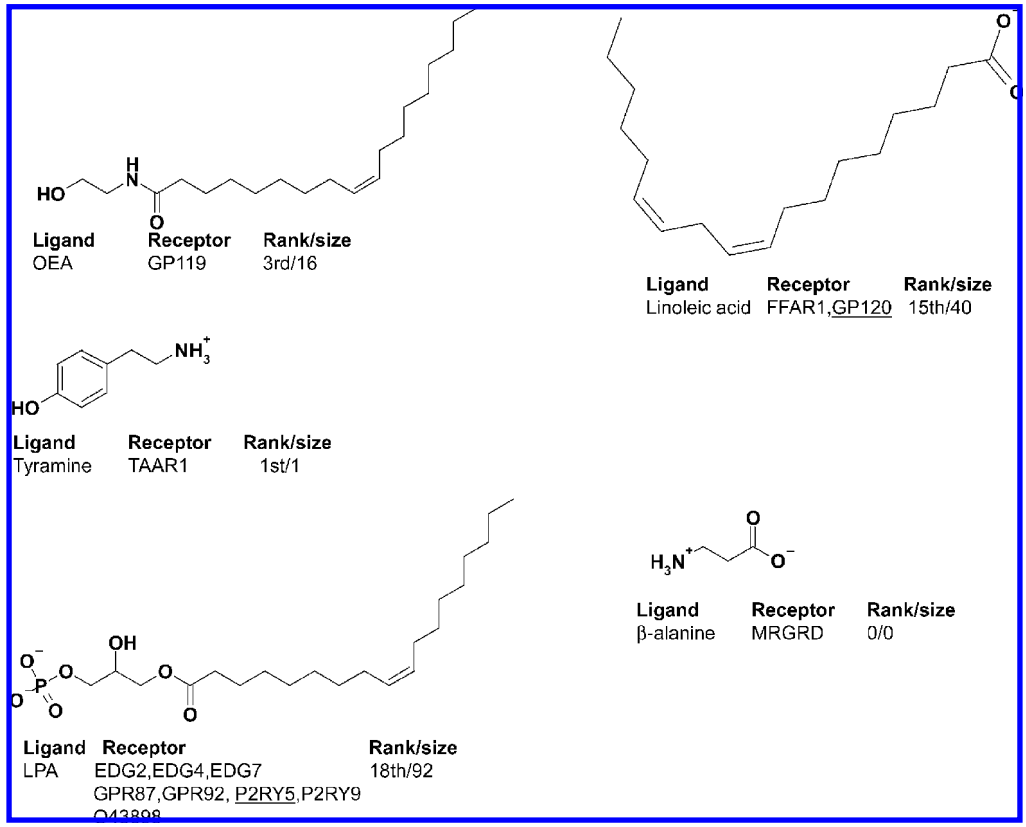

Figure 8. Examples of receptor prediction for various endogenous ligands of the external test set 1, using the MACCS-RF model trained on Data set 1. The name of the true GPCR is indicated along with its rank and the size of the corresponding receptor hit list. In case of binding to multiple receptors, the highest-ranked true receptor is underlined.

is found in various GPCR activity classes (PRED23 resembles much more adenosine and biogenic amine receptor ligands than known OXYR ligands; Figure 7), the corresponding hit list is larger since it contains many receptors of the corresponding GPCR clusters and the rank of the true receptor is low.

In receptor prediction mode, local models again outperform the global model (Figure 6 C,D). Whatever the external test set, the global model disappointingly produces unspecific and large hit lists (above 100 receptors) with very high average ranks of the true receptor (Figure 6C). Therefore, it cannot be used to profile a particular ligand against a large panel of GPCRs. Local models exhibit a much better performance. The MACCS/RF combination is again excellent for the test set of endogenous ligands. Figure 8 shows a few representative results obtained with this model (see all results in Supporting Information Table

S8). GPR119, a recently deorphanized receptor for oleylethanolamide (OEA),[76] is ranked third among the 16 predicted OEA receptors, although no GPR119 ligands are stored in the training Data set 1. The PLFP is particularly well suited for this example since OEA share many structural features with anandamide (AEA), an endogenous ligand for cannabinoid receptors,[77] while GPR119 and cannabinoid receptor cavities are also quite similar.[26] The trace amine TAAR1 receptor is retrieved as the only putative receptor for tyramine (true TAAR1 ligand) although more than 500 known GPCR ligands (mostly for biogenic amine receptors) in the training set share the tyramine substructure. The case of the fatty linoleic acid, a ligand for unrelated FFAR1 and GP120 entries, is more difficult. Both GPCRs have a very low sequence identity (13%) and are located in distant branches of our cavity-biased tree. No GPR120 ligands and only 3 FFAR1 ligands have been used to train the classification model. Nevertheless, the two receptors are selected in the hit list, GP120 being the highest ranked (15th, Figure 8) but after false positives (e.g., prostaglandin receptors). If the test ligand is very permissive for a wide array of receptors spread over the entire GPCR space (e.g., lysophosphatidic acid), the receptor hit list is very large with many false negatives (S1P, prostaglandin and leukotriene receptors). Interestingly, no predictions are made in the absence of data (e.g., no receptor list for beta-alanine) which perfectly illustrates the fact that QSAR models always have a limited applicability domain.

When applied to the set of synthetic ligands, the same trend as previously reported in ligand prediction mode is observed: the recall drops significantly. The best compromise is achieved with the DistFP/SVM model (Figure 6D). This classification model shows a good recall value (0.64) and can still be applied for an experimental validation since it would require testing a particular compound on an average list of 30 receptors.

Several conclusions can be drawn from this first series of predictions: i) classification models perform better when applied to a target-ligand space already covered by the training set, ii) in the absence of data, no prediction could be made, iii) local models outperform a global model just because they better address local properties of the protein−ligand space under investigation, iv) the best cross-validated model is not necessarily the most predictive one for external test sets, and v) predicting ligands for a given receptor is easier than predicting receptors for a given ligand.

The first four conclusions may not be surprising, but the corresponding issues have not been fully addressed in previous studies using PLFPs.[17,19,20] The last conclusion probably results from a bias in defining decoys in the training Data set 1. Selecting decoys is an intense matter of debate in the field of virtual screening,[36,78] and no optimal solution has been found yet. In the current implementation of PLFPs, the variability of the ligand block is significantly higher than that of the receptor block. Decoys here should describe GPCR-ligand pairs with a very low probability of occurrence. Since ligand space is much larger than GPCR space, it was tempting to choose GPCR-ligand decoys by varying the ligand part of the PLFP. However, the chemical space from which decoy ligands are chosen (commercially available druglike compounds) is different from that covered by true actives (MDDR GPCR ligands). We therefore generated a second data set (Data set 2) in which decoy ligands were chosen exactly in the same chemical space as the true actives.

Classification and Prediction Models from Data Set 2: Influence of the Decoys Set. Data set 2 differs from the first one in the way presumably inactive pairs had been generated. The aim of the current study is not to present a novel way of selecting decoys but just to pinpoint the influence of decoys selection on results, notably when the classification method is switched from ligand-prediction to receptor-prediction mode. In data set 2, all 21050 MDDR ligands are considered in association with their cognate receptors as true actives (32118 pairs in total). The 202019 inactive pairs were chosen by linking the same MDDR ligands with GPCR entries too far away, according to our CavFP cavity fingerprint, from their true receptors (see Computational Methods). A minimum Euclidean distance of 7.0 between two GPCR entries was chosen since it allows a perfect separation of GPCRs across the 22 clusters.

Cross-validation Models. Only local models applied to Data set 2 will be discussed here because they unambiguously showed a much better predictive performance than global models on previous external data sets. Most local models show excellent performances (recall, precision, F-measure; Figure 9A). When compared to values obtained for Data set 1, it appears clearly that modifying the way GPCR-ligand decoys have been constructed does not affect the accuracy of the models (compare Figure 4B with Figure 9A). However, since the performance of the cross-validated model was not a good indicator of its true predictivity when applied to external test sets, we applied the same external validation as before on both sets of endogenous and synthetic GPCR ligands.

Predicting GPCR-Ligand Pairs. In ligand prediction mode local models are generally less accurate than those previously obtained from Data set 1. Recall values are usually lower (0.6−0.7 instead of 0.7−0.85), hit lists larger, and rank of true ligands higher (compare Figure 6B with Figure 9B). The MACCS/RF model which is again the best method for predicting ligands is however inferior to the same model derived from Data set 1.

As expected however, predictions in receptor mode (ligand descriptor is kept constant) are now better (compare Figure 6D with Figure 9C). The MACCS/RF model for example is very satisfactory; it generates receptor lists of reasonable size (ca. 20 receptors) with an usually low rank of the fist-ranked true GPCR (3−5) and an acceptable recall value (0.4−0.66), regarding the complexity of the request. In 23 out of 35 test cases (external set 1), the true receptor is recovered in the hit list, quite often in the first position (see Supporting Information Table S9 for a summary of all results). For example, OXGR1 is indeed ranked first as putative receptor for alpha-ketoglutaric acid (Figure 10). Interestingly, the receptor list is quite short (5 entries) and notably encloses receptors for chemically similar ligands like succinic acid (SUCR1)[79] and short chain fatty acids (FFAR2, FFAR3).[80]

The model is usually quite selective for the true receptors when several subtypes are equally potent in ligand binding. Hence, all four adenosine receptors are equally ranked at first place as receptors of adenosine with very few false positives. In our approach, any receptor which is given a probability higher than 0.5 is selected as hit. Our data however suggest than increasing the probability threshold to a higher value (e.g., 0.8) would considerably decrease the size of the receptor list with limited influence on the recall.

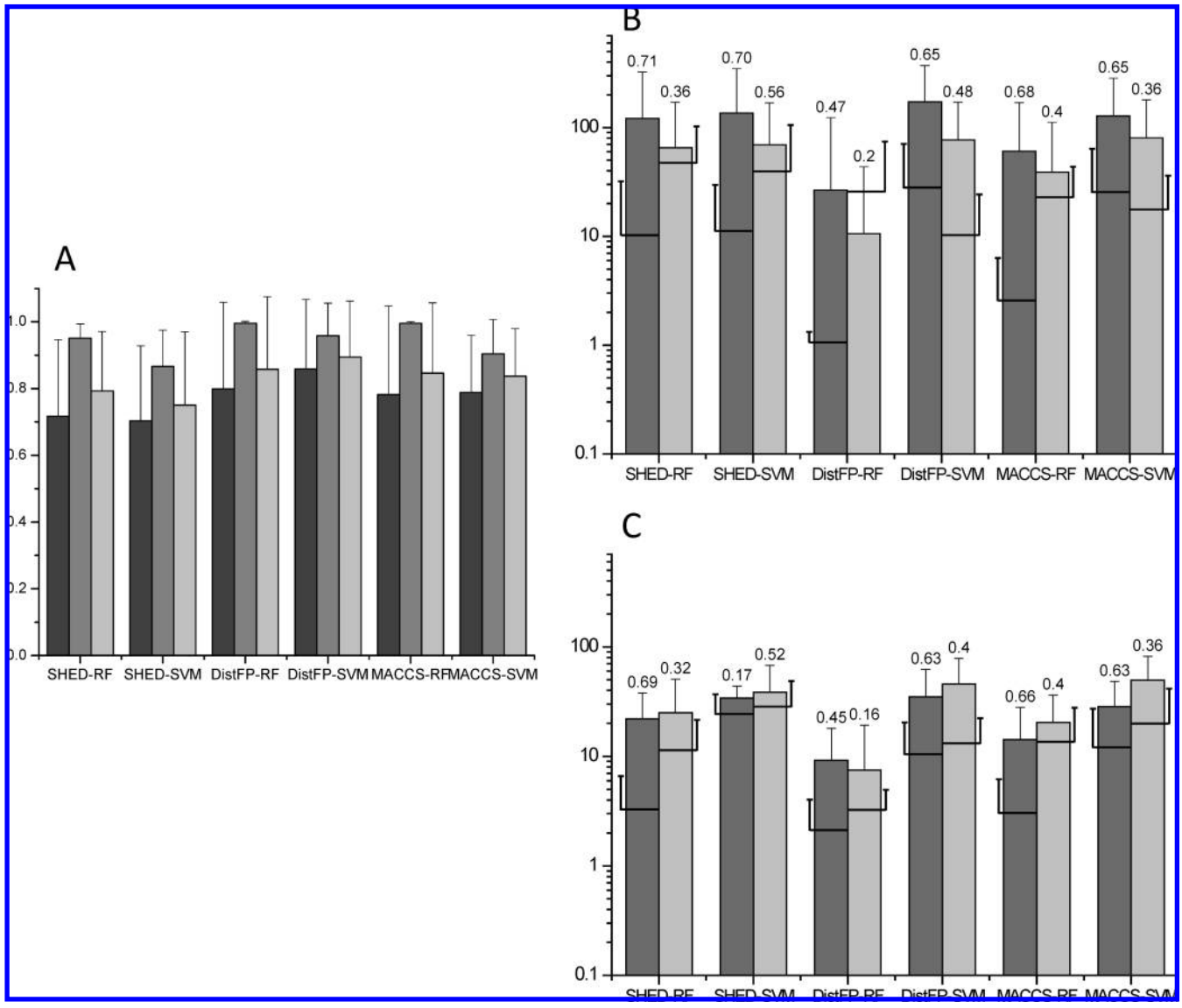

Figure 9. A) Performance of 19 local models (panel A) in classifying and predicting GPCR-ligand pairs from Data set 2. The mean value and standard deviation is given for all local models outputting a probability of binary association higher than 0.5. Criteria used to judge the performance are recall (dark gray bars), precision (gray bars), and F-measure (light gray bars). **B,C)** Predicting possible protein−ligand pairs for two external test sets (test set 1 of 35 endogenous GPCR ligands, dark gray bars; test set 2 of 25 synthetic GPCR ligands, light gray bars) with 19 local models trained on Data set 2 (see Methods). The size of the ligand (panels B) or the receptor hit list (panel C) is shown by a bar, and the rank of the true ligand/receptor in the hit list is displayed by a horizontal bar-crossing line. Vertical lines starting from the bar and the bar-crossing line indicate the standard deviation of the hit list size and of the true hit rank, respectively. The number on the top of the bar is the recall of the prediction on the entire test set.

For example, 41 GPCRs of the biogenic amine receptor family are predicted as putative receptor of adrenaline ($p >$ 0.5). Using a probability threshold of 0.8, only 12 entries are selected, out of which the 8 true adrenaline receptors are present. It is however quite difficult to customize such a threshold for all receptors since many entries are described by a limited number of ligands. In case the receptor prediction produces a too large hit list, it might however make sense to prioritize the top-ranked entries for experimental validation.

When profiling synthetic ligands, the receptor lists are generally larger and recall values inferior to that observed for endogenous ligands (Figure 9C), but the true receptor is found in 10 out of 25 possible cases. When the ligand is very specific for a given GPCR space, the receptor list remains short. An illustrative example is the profiling of the selective MC4R antagonist (PRED14;[49] Figure 10) which returns the true receptor at first rank of a rather short receptor list (Figure 10). Profiling a more permissive ligand (e.g., PRED11;[53] Figure 10) produces larger receptor lists, but the true receptors are still ranked high, possibly with false positives but also false negatives (the top-ranked DRD3 receptor being is likely a true receptor of this ligand).[53]

A general trend which is observed, irrespective of the data set, is that predictions in both modes (receptor mode, ligand mode) are very often more accurate when applied to the set of endogenous ligands which are indeed of lower complexity (molecular weight, heavy atom counts) than the set of synthetic ligands. This opens the door to scaffold hopping to a particular GPCR subspace. The second trend is that predicting ligands is easier than predicting receptors. The main explanation for this observation is that the ligand part of the PLFP is probably more variable than the receptor part. Alternative receptor descriptions still focusing on the trans-

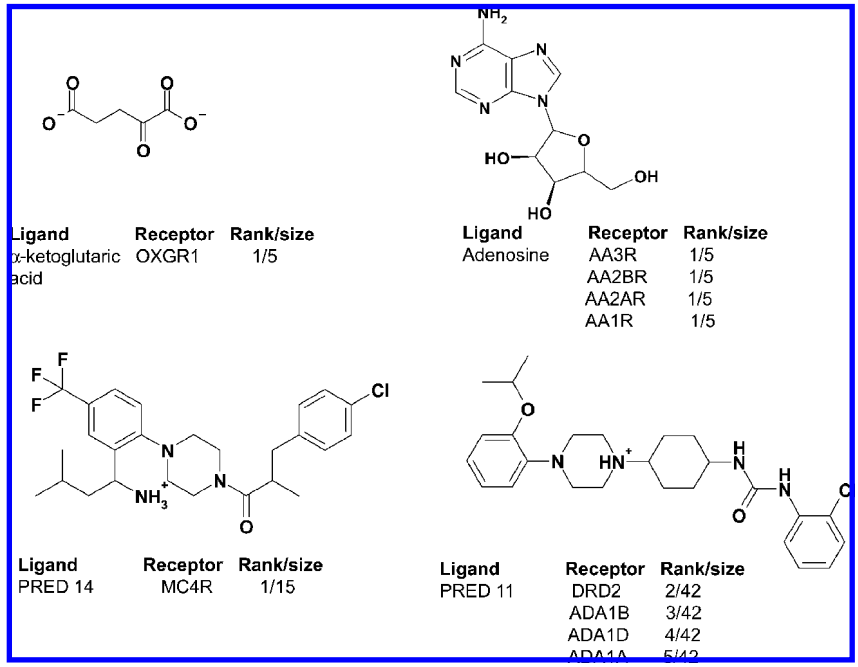

Figure 10. Examples of receptor prediction for four ligands of the external test sets 1 and 2, using the MACCS-RF model trained on Data set 2. The name of the true GPCR is indicated along with its rank and the size of the corresponding receptor hit list.

membrane binding site (e.g., receptor-derived pharmacophore fingerprints)[81] will be investigated in the near future to address this issue. Using the inner products of vectors describing both cavities and ligands[19] instead of concatenating cavity and ligand vectors is another possibility to increase the weight put on cavities arising from an homogeneous family of similar receptors.

Comparison of PFLPs with Other Protein−Ligand Fingerprints. The current study and PFLP descriptor differ significantly from related chemogenomic approaches already presented by three other groups. In the proteochemometrics (PCM) approach developed by Wikberg et al.,[17,18] proteins are encoded by global descriptors based on the amino acid sequence (occurrence of specific amino acids or principal components of residue properties) or the topology of the corresponding active site (e.g., binding site surface area). Local structure-based descriptors of enzyme clefts were recently introduced in PCM[17] by registering the local amino acid neighborhood of any protein residue in a binding pocket, but the description is still based on the amino acid sequence and not precise pharmacophoric 3-D information as in PLFPs. Similarly, Bock et al.[20] describe the target by global features (surface tension, isoelectric point, accessible surface area) of the full amino acid sequence without focusing on the binding site. It presents the advantage to be independent of the prior knowledge of the ligand-binding site but does not take into account property selection pressure[82] along the amino acid sequence whether a residue is part of a ligand-binding site or not. Our approach presents the advantage to be highly focused on the protein−ligand interface by selecting only ligands binding to the transmembrane cavity and describing the latter cavities by pharmacophoric descriptors. It is closer to the binding pocket kernel introduced by Jacob et al. for the transmembrane cavity of GPCRs.[19] The main difference between the current study and Jacob's report lies in the data sets of ligands and targets used for learning. Whereas Jacob et al. learned from 4051 GPCR-ligand interactions inferred from the GLIDA GPCR ligand database[23] and a collection of 80 human GPCRs, our data set is significantly more exhaustive in terms of protein−ligand pairs (32 118 distinct pairs) and biological space coverage (160 sequences).

Comparing PLFPs with Ligand Fingerprints. Does encoding both protein cavity and ligand properties in a same fingerprint provide some advantages over pure ligand-based similarity search approaches[6−10] relying on ligand properties only? To address this question, we compared PLFPs with the corresponding ligand fingerprints (SHED, DistFP, MACCS) in their capacity to recover the true receptor (receptor prediction mode) or the true ligand (ligand prediction mode) of ligands/receptors present in both external test sets. We should recall here that the purpose of this experiment is not to compare PLFPs with state-of-the-art ligand descriptors (e.g., circular fingerprints)[68] but just to evaluate the exact influence of the receptor cavity block in the fingerprint.

In the first approach, classification models were computed for each GPCR activity class (160 in total). Results were disappointing (data not shown) mainly because of the lack of really diverse ligands for numerous GPCR entries. We then decided to investigate a simple nearest-neighbor approach using a Tanimoto coefficient on MACCS keys as metric. In receptor prediction mode, all 21050 GPCR ligands of the training set were ranked by decreasing similarity to the ligand of the external test set, and corresponding receptors were ranked accordingly. In ligand prediction mode, each external ligand with its own set of decoys ("query") was compared to all known ligands of the corresponding GPCR in the training set, and the query molecules were then ranked by decreasing similarity score to any known GPCR ligand of that receptor (Supporting Information Figure 1). Since MACCS keys gave the best results among the three investigated descriptors, we will discuss comparative results between ligand-based and protein−ligand fingerprints using MACCS keys only.

Table 1. Mean Rank of the True Receptor (Receptor Prediction Mode) and of the True Ligand (Ligand Prediction Mode) in Two External Validations of Classification Models Applied to Protein−Ligand and Ligand Fingerprints

	receptor prediction mode		ligand prediction mode	
descriptor	Set1[a]	Set2[b]	Set1	Set2
ligand[c]	11.3	45.6	1.1	15.3
PLFP[d]	4.8	25.7	1.2	3.6

[a] Set of 35 nonpeptidic endogenous ligands targeting 88 GPCRs. [b] Set of 25 synthetic ligands targeting 28 GPCRs. [c] Ligand fingerprint: 166-bit public MACCS keys. [d] MACCS-CavFP fingerprint using RF (receptor prediction mode: sets 1 and 2, ligand prediction mode:set 1) or SVM (ligand prediction mode: set 2) as machine learning classifier.

In receptor prediction mode, the PLFP outperformed the corresponding MACCS keys in ranking as high as possible the true receptor of 60 ligands from both external test sets (Table 1). In ligand prediction mode, the PLFP has a performance similar to that of the MACCS key for simple ligands (test set 1 of endogenous ligands) but much better as far as the complexity of the molecules increases (test set 2 of synthetic ligands). A significant difference between PLFPs and simple ligand fingerprints is their much wider applicability. A pure ligand-based similarity search is restricted in our case to 160 activity classes (receptors) for which ligands are reported. Using PLFPs, only receptors belonging to 3 GPCR subclasses (Adhesion, MAS, SRBs)[26] cannot be considered. Predictions can thus be extended to a total of 320 possible activity classes including orphan receptors.

CONCLUSIONS

Protein−ligand fingerprints (PLFPs) represent a novel way of encoding structural information on both ligands and their corresponding binding sites in order to directly mine chemogenomic space. Encoding target profiles by NB modeling on multiple activity classes into "Bayes affinity fingerprints" was already shown to be superior to conventional ligand-based similarity searches in comparing bioactive ligands.[6] The herein presented PLFP descriptor goes one step further by adding to the same vector crucial structural details of protein−ligand interfaces. The current study unambiguously demonstrates that protein−ligand fingerprints outperform the corresponding ligand fingerprints in predicting either putative ligands for a known target or putative targets for a known ligand. In the current implementation, focusing on a protein family, predicting ligands (ligand screening), is significantly easier than predicting targets (target profiling). Further studies are still required to find the proper balance between ligand and protein description using specifically designed kernels.[83] PLFPs are probably better suited to browse the full chemogenomic space than a target family subspace as described in the current study, for the simple reason that ligand space is several orders of magnitude more diverse than target space. The broader the target space, the more variable cavity descriptors should be. It however requires a simple, generic, and size-independent description of protein binding sites which is still missing today. The current application to the universe of GPCRs shows that PLFPs are in most cases superior to the corresponding ligand-derived descriptors and most importantly applicable to a much larger chemogenomic space. It also demonstrates that numerous checks on various parameters (critical evaluation of ligand and target space coverage, machine learning algorithm, selection of decoys, true validation with external test sets, and comparative evaluation of global versus local models) are mandatory to optimize the peak performance of QSAR models according to the screening scenario.

ACKNOWLEDGMENT

This work was supported by a grant from the French Ministry of Research to N.W. We thank Dr. Esther Kellenberger, Dr. Jérome Hert, and Dr. Nathan Brown for helpful discussions and critical reading of the manuscript. The full data set of receptor-annotated ligands is available upon request to the authors and certification of a valid MDDR licence.

Supporting Information Available: Correspondence table between MDDR activity classes and GPCR entry names (Table S1), parameters for Filter 2.0.1 (OpenEye Scientific Software) (Table S2), parameters for Standardizer (ChemAxon Ltd.) (Table S3), list of 35 endogenous ligands (external test set 1) (Table S4), list of 25 synthetic ligands (external test set 2) (Table S5), cavity fingerprint (CavFP) definition (Table S6), property mapping for computing SHED descriptors and distributions (Table S7), rank of the true ligand/receptor hit and size of the ligand/receptor hit lists in predicting ligands/receptors for both external test sets (set 1 of 35 endogenous GPCR ligands, set 2 of 25 synthetic GPCR ligands) for global and local models generated from Data set 1 (Table S8), rank of the true ligand/receptor hit and size of the ligand/receptor hit lists in predicting ligands/receptors for both external test sets (set 1 of 35 endogenous GPCR ligands, set 2 of 25 synthetic GPCR ligands) for local models generated from Data set 2 (Table S9), and nearest-neighbor similarity searches from ligand descriptors (Figure S1). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Bredel, M.; Jacoby, E. Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* **2004**, *5*, 262–275.
(2) Harris, C. J.; Stevens, A. P. Chemogenomics: structuring the drug discovery process to gene families. *Drug Discovery Today* **2006**, *11*, 880–888.
(3) Caron, P. R.; Mullican, M. D.; Mashal, R. D.; Wilson, K. P.; Su, M. S.; Murcko, M. A. Chemogenomic approaches to drug discovery. *Curr. Opin. Chem. Biol.* **2001**, *5*, 464–470.
(4) Krejsa, C. M.; Horvath, D.; Rogalski, S. L.; Penzotti, J. E.; Mao, B.; Barbosa, F.; Migeon, J. C. Predicting ADME properties and side effects: the BioPrint approach. *Curr. Opin. Drug Discovery Dev.* **2003**, *6*, 470–480.
(5) Rognan, D. Chemogenomic approaches to rational drug design. *Br. J. Pharmacol.* **2007**, *152*, 38–52.
(6) Bender, A.; Jenkins, J. L.; Glick, M.; Deng, Z.; Nettles, J. H.; Davies, J. W. "Bayes affinity fingerprints" improve retrieval rates in virtual screening and define orthogonal bioactivity space: when are multitarget drugs a feasible concept. *J. Chem. Inf. Model.* **2006**, *46*, 2445–2456.
(7) Nettles, J. H.; Jenkins, J. L.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M. Bridging chemical and biological space: "target fishing" using 2D and 3D molecular descriptors. *J. Med. Chem.* **2006**, *49*, 6802–6810.
(8) Nidhi; Glick, M.; Davies, J. W.; Jenkins, J. L. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J. Chem. Inf. Model.* **2006**, *46*, 1124–1133.
(9) Mestres, J.; Gregori-Puigjane, E.; Valverde, S.; Sole, R. V. Data completeness-the Achilles heel of drug-target networks. *Nat. Biotechnol.* **2008**, *26*, 983–984.

(10) 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.
(11) 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.
(12) Schmitt, S.; Kuhn, D.; Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* **2002**, *323*, 387–406.
(13) Gold, N. D.; Jackson, R. M. Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. *J. Mol. Biol.* **2006**, *355*, 1112–1124.
(14) Schalon, C.; Surgand, J. S.; Kellenberger, E.; Rognan, D. A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins* **2008**, *71*, 1755–1778.
(15) Klabunde, T. Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. *Br. J. Pharmacol.* **2007**, *152*, 5–7.
(16) Erhan, D.; L'Heureux, P. J.; Yue, S. Y.; Bengio, Y. Collaborative filtering on a family of biological targets. *J. Chem. Inf. Model.* **2006**, *46*, 626–635.
(17) Strombergsson, H.; Daniluk, P.; Kryshtafovych, A.; Fidelis, K.; Wikberg, J. E.; Kleywegt, G. J.; Hvidsten, T. R. Interaction Model Based on Local Protein Substructures Generalizes to the Entire Structural Enzyme-Ligand Space. *J. Chem. Inf. Model.* **2008**, *48*, 2278–2288.
(18) Lapinsh, M.; Prusis, P.; Gutcaits, A.; Lundstedt, T.; Wikberg, J. E. Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions. *Biochim. Biophys. Acta* **2001**, *1525*, 180–190.
(19) Jacob, L.; Hoffmann, B.; Stoven, V.; Vert, J. P. Virtual screening of GPCRs: an in silico chemogenomics approach. *BMC Bioinf.* **2008**, *9*, 363.
(20) Bock, J. R.; Gough, D. A. Virtual screen for ligands of orphan G protein-coupled receptors. *J. Chem. Inf. Model* **2005**, *45*, 1402–1414.
(21) Lagerstrom, M. C.; Schioth, H. B. Structural diversity of G protein-coupled receptors and significance for drug discovery. *Nat. Rev. Drug Discovery* **2008**, *7*, 339–357.
(22) Jensen, N. H.; Roth, B. L. Massively parallel screening of the receptorome. *Comb. Chem. High Throughput Screening* **2008**, *11*, 420–426.
(23) Okuno, Y.; Tamon, A.; Yabuuchi, H.; Niijima, S.; Minowa, Y.; Tonomura, K.; Kunimoto, R.; Feng, C. GLIDA: GPCR-ligand database for chemical genomics drug discovery--database and tools update. *Nucleic Acids Res.* **2008**, *36*, D907–912.
(24) Klabunde, T.; Jager, R. Chemogenomics approaches to G-protein coupled receptor lead finding. *Ernst Schering Res. Found. Workshop* **2006**, 31–46.
(25) Frimurer, T. M.; Ulven, T.; Elling, C. E.; Gerlach, L. O.; Kostenis, E.; Hogberg, T. A physicogenetic method to assign ligand-binding relationships between 7TM receptors. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3707–3712.
(26) 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.
(27) Bjarnadottir, T. K.; Gloriam, D. E.; Hellstrand, S. H.; Kristiansson, H.; Fredriksson, R.; Schioth, H. B. Comprehensive repertoire and phylogenetic analysis of the G protein-coupled receptors in human and mouse. *Genomics* **2006**, *88*, 263–273.
(28) Warne, T.; Serrano-Vega, M. J.; Baker, J. G.; Moukhametzianov, R.; Edwards, P. C.; Henderson, R.; Leslie, A. G.; Tate, C. G.; Schertler, G. F. Structure of a beta1-adrenergic G-protein-coupled receptor. *Nature* **2008**, *454*, 486–491.
(29) Symyx Technologies, Inc., Santa Clara, CA.
(30) OpenEye Scientific Software, Santa Fe, NM 87507.
(31) ChemAxon Kft., Budapest 1037, Hungary.
(32) SciTegic Inc., San Diego, CA 92123-1365, U.S.A.
(33) Johnson, M. E.; Moore, L. M.; Ylvisaker, D. Minimax and maximin distance designs. *J. Statist. Plann. Infer.* **1990**, *26*, 131–148.
(34) Chemical Computing Group Inc., Montreal, Quebec, Canada.
(35) http://bioinfo-pharma.u-strasbg.fr/bioinfo (accessed Feb 2009).
(36) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.
(37) Yang, L.; Zhou, C.; Guo, L.; Morriello, G.; Butora, G.; Pasternak, A.; Parsons, W. H.; Mills, S. G.; MacCoss, M.; Vicario, P. P.; Zweerink, H.; Ayala, J. M.; Goyal, S.; Hanlon, W. A.; Cascieri, M. A.; Springer, M. S. Discovery of 3,5-bis(trifluoromethyl)benzyl l-arylglycinamide based potent CCR2 antagonists. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 3735–3739.
(38) Di Fabio, R.; Giovannini, R.; Bertani, B.; Borriello, M.; Bozzoli, A.; Donati, D.; Falchi, A.; Ghirlanda, D.; Leslie, C. P.; Pecunioso, A.; Rumboldt, G.; Spada, S. Synthesis and SAR of substituted tetrahydrocarbazole derivatives as new NPY-1 antagonists. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 1749–1752.
(39) Xie, Y. F.; Sircar, I.; Lake, K.; Komandla, M.; Ligsay, K.; Li, J.; Xu, K.; Parise, J.; Schneider, L.; Huang, D.; Liu, J.; Sakurai, N.; Barbosa, M.; Jack, R. Identification of novel series of human CCR1 antagonists. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 2215–2221.
(40) Micheli, F.; Bertani, B.; Bozzoli, A.; Crippa, L.; Cavanni, P.; Di Fabio, R.; Donati, D.; Marzorati, P.; Merlo, G.; Paio, A.; Perugini, L.; Zarantonello, P. Phenylethynyl-pyrrolo[1,2-a]pyrazine: A new potent and selective tool in the mGluR5 antagonists arena. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 1804–1809.
(41) Feng, D.-M.; DiPardo, R. M.; Wai, J. M.; Chang, R. K.; Di Marco, C. N.; Murphy, K. L.; Ransom, R. W.; Reiss, D. R.; Tang, C.; Prueksaritanont, T.; Pettibone, D. J.; Bock, M. G.; Kuduk, S. D. A new class of bradykinin B1 receptor antagonists with high oral bioavailability and minimal PXR activity. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 682–687.
(42) Erickson, S. D.; Banner, B.; Berthel, S.; Conde-Knape, K.; Falcioni, F.; Hakimi, I.; Hennessy, B.; Kester, R. F.; Kim, K.; Ma, C.; McComas, W.; Mennona, F.; Mischke, S.; Orzechowski, L.; Qian, Y.; Salari, H.; Tengi, J.; Thakkar, K.; Taub, R.; Tilley, J. W.; Wang, H. Potent, selective MCH-1 receptor antagonists. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 1402–1406.
(43) Bergman, J. M.; Roecker, A. J.; Mercer, S. P.; Bednar, R. A.; Reiss, D. R.; Ransom, R. W.; Meacham Harrell, C.; Pettibone, D. J.; Lemaire, W.; Murphy, K. L.; Li, C.; Prueksaritanont, T.; Winrow, C. J.; Renger, J. J.; Koblan, K. S.; Hartman, G. D.; Coleman, P. J. Proline bis-amides as potent dual orexin receptor antagonists. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 1425–1430.
(44) Lewis, L. M.; Sheffler, D.; Williams, R.; Bridges, T. M.; Kennedy, J. P.; Brogan, J. T.; Mulder, M. J.; Williams, L.; Nalywajko, N. T.; Niswender, C. M.; Weaver, C. D.; Conn, P. J.; Lindsley, C. W. Synthesis and SAR of selective muscarinic acetylcholine receptor subtype 1 (M1 mAChR) antagonists. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 885–890.
(45) Yoon, T.; De Lombaert, S.; Brodbeck, R.; Gulianello, M.; Chandrasekhar, J.; Horvath, R. F.; Ge, P.; Kershaw, M. T.; Krause, J. E.; Kehne, J.; Hoffman, D.; Doller, D.; Hodgetts, K. J. The design, synthesis and structure-activity relationships of 1-aryl-4-aminoalkyl-isoquinolines: A novel series of CRF-1 receptor antagonists. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 891–896.
(46) Micheli, F.; Bonanomi, G.; Braggio, S.; Capelli, A. M.; Celestini, P.; Damiani, F.; Fabio, R. D.; Donati, D.; Gagliardi, S.; Gentile, G.; Hamprecht, D.; Petrone, M.; Radaelli, S.; Tedesco, G.; Terreni, S.; Worby, A.; Heidbreder, C. New fused benzazepine as selective D3 receptor antagonists. Synthesis and biological evaluation. Part one: [h]-fused tricyclic systems. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 901–907.
(47) Troxler, T.; Enz, A.; Hoyer, D.; Langenegger, D.; Neumann, P.; Pfäffli, P.; Schoeffter, P.; Hurth, K. Ergoline derivatives as highly potent and selective antagonists at the somatostatin sst1 receptor. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 979–982.
(48) Beck, H. P.; Kohn, T.; Rubenstein, S.; Hedberg, C.; Schwandner, R.; Hasslinger, K.; Dai, K.; Li, C.; Liang, L.; Wesche, H.; Frank, B.; An, S.; Wickramasinghe, D.; Jaen, J.; Medina, J.; Hungate, R.; Shen, W. Discovery of potent LPA2 (EDG4) antagonists as potential anticancer agents. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 1037–1041.
(49) Tran, J. A.; Chen, C. W.; Tucci, F. C.; Jiang, W.; Fleck, B. A.; Chen, C. Syntheses of tetrahydrothiophenes and tetrahydrofurans and studies of their derivatives as melanocortin-4 receptor ligands. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 1124–1130.
(50) Li, G.; Stamford, A. W.; Huang, Y.; Cheng, K.-C.; Cook, J.; Farley, C.; Gao, J.; Ghibaudi, L.; Greenlee, W. J.; Guzzi, M.; van Heek, M.; Hwa, J. J.; Kelly, J.; Mullins, D.; Parker, E. M.; Wainhaus, S.; Zhang, X. Discovery of novel orally active ureido NPY Y5 receptor antagonists. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 1146–1150.
(51) Foloppe, N.; Allen, N. H.; Bentley, C. H.; Brooks, T. D.; Kennett, G.; Knight, A. R.; Leonardi, S.; Misra, A.; Monck, N. J. T.; Sellwood, D. M. Discovery of a novel class of selective human CB1 inverse agonists. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 1199–1206.
(52) Du, X.; Chen, X.; Mihalic, J. T.; Deignan, J.; Duquette, J.; Li, A.-R.; Lemon, B.; Ma, J.; Miao, S.; Ebsworth, K.; Sullivan, T. J.; Tonn, G.; Collins, T. L.; Medina, J. C. Design and optimization of imidazole derivatives as potent CXCR3 antagonists. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 608–613.
(53) Chiu, G.; Li, S.; Connolly, P. J.; Pulito, V.; Liu, J.; Middleton, S. A. (Phenylpiperazinyl)cyclohexylureas: Discovery of [alpha]1a/1d-selective adrenergic receptor antagonists for the treatment of benign prostatic hyperplasia/lower urinary tract symptoms (BPH/LUTS). *Bioorg. Med. Chem. Lett.* **2008**, *18*, 640–644.
(54) Seong, C. M.; Park, W. K.; Park, C. M.; Kong, J. Y.; Park, N. S. Discovery of 3-aryl-3-methyl-1H-quinoline-2,4-diones as a new class of selective 5-HT6 receptor antagonists. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 738–743.

(55) Nguyen, D. N.; Paone, D. V.; Shaw, A. W.; Burgey, C. S.; Mosser, S. D.; Johnston, V.; Salvatore, C. A.; Leonard, Y. M.; Miller-Stein, C. M.; Kane, S. A.; Koblan, K. S.; Vacca, J. P.; Graham, S. L.; Williams, T. M. Calcitonin gene-related peptide (CGRP) receptor antagonists: Investigations of a pyridinone template. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 755–758.

(56) Xiao, Y.; Araldi, G. L.; Zhao, Z.; Reddy, A.; Karra, S.; Brugger, N.; Fischer, D.; Palmer, E.; Bao, B.; McKenna, S. D. Synthesis and evaluation of a [gamma]-lactam as a highly selective EP2 and EP4 receptor agonist. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 821–824.

(57) Ly, K. S.; Letavic, M. A.; Keith, J. M.; Miller, J. M.; Stocking, E. M.; Barbier, A. J.; Bonaventure, P.; Lord, B.; Jiang, X.; Boggs, J. D.; Dvorak, L.; Miller, K. L.; Nepomuceno, D.; Wilson, S. J.; Carruthers, N. I. Synthesis and biological activity of piperazine and diazepane amides that are histamine H3 antagonists and serotonin reuptake inhibitors. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 39–43.

(58) Liddle, J.; Allen, M. J.; Borthwick, A. D.; Brooks, D. P.; Davies, D. E.; Edwards, R. M.; Exall, A. M.; Hamlett, C.; Irving, W. R.; Mason, A. M.; McCafferty, G. P.; Nerozzi, F.; Peace, S.; Philp, J.; Pollard, D.; Pullen, M. A.; Shabbir, S. S.; Sollis, S. L.; Westfall, T. D.; Woollard, P. M.; Wu, C.; Hickey, D. M. B. The discovery of GSK221149A: A potent and selective oxytocin antagonist. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 90–94.

(59) Skinner, P. J.; Cherrier, M. C.; Webb, P. J.; Sage, C. R.; Dang, H. T.; Pride, C. C.; Chen, R.; Tamura, S. Y.; Richman, J. G.; Connolly, D. T.; Semple, G. 3-Nitro-4-amino benzoic acids and 6-amino nicotinic acids are highly selective agonists of GPR109b. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 6619–6622.

(60) Guba, W.; Green, L. G.; Martin, R. E.; Roche, O.; Kratochwil, N.; Mauser, H.; Bissantz, C.; Christ, A.; Stahl, M. From Astemizole to a Novel Hit Series of Small-Molecule Somatostatin 5 Receptor Antagonists via GPCR Affinity Profiling. *J. Med. Chem.* **2007**, *50*, 6295–6298.

(61) Sneath, P. H.; Sokal, R. R. Numerical taxonomy. *Nature* **1962**, *193*, 855–860.

(62) Felsendtein, J. PHYLIP - Phylogeny Inference Package (version 3.2). *Cladistics* **1989**, *5*, 164–166.

(63) Gregori-Puigjane, E.; Mestres, J. SHED: Shannon entropy descriptors from topological feature distributions. *J. Chem. Inf. Model.* **2006**, *46*, 1615–1622.

(64) TRIPOS, Assoc., Inc., St. Louis, MO.

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

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

(67) Chang, C. C.; Lin, C. J. LIBSVM: a library for support vector machines (software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm), 2001.

(68) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.

(69) Schuffenhauer, A.; Zimmermann, J.; Stoop, R.; van der Vyver, J. J.; Lecchini, S.; Jacoby, E. An ontology for pharmaceutical ligands and its application for in silico screening and library design. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 947–955.

(70) Tyndall, J. D. A.; Pfeiffer, B.; Abbenante, G.; Fairly, D. P. Over 100 Peptide-Activated G Protein-Coupled Receptors Recognize Ligands With Turn Structure. *Chem. Rev.* **2004**, *105*, 793–826.

(71) Cleves, A. E.; Jain, A. N. Effects of inductive bias on computational evaluations of ligand-based modeling and on drug discovery. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 147–59.

(72) Soudijn, W.; van Wijngaarden, I.; Ijzerman, A. P. Nicotinic acid receptor subtypes and their ligands. *Med. Res. Rev.* **2007**, *27*, 417–433.

(73) Fredriksson, R.; Lagerstrom, M. C.; Lundin, L. G.; Schioth, H. B. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol. Pharmacol.* **2003**, *63*, 1256–1272.

(74) Gregori-Puigjane, E.; Mestres, J. A ligand-based approach to mining the chemogenomic space of drugs. *Comb. Chem. High Throughput Screening* **2008**, *11*, 669–676.

(75) Peltason, L.; Bajorath, J. SAR index: quantifying the nature of structure-activity relationships. *J. Med. Chem.* **2007**, *50*, 5571–5578.

(76) Overton, H. A.; Babbs, A. J.; Doel, S. M.; Fyfe, M. C.; Gardner, L. S.; Griffin, G.; Jackson, H. C.; Procter, M. J.; Rasamison, C. M.; Tang-Christensen, M.; Widdowson, P. S.; Williams, G. M.; Reynet, C. Deorphanization of a G protein-coupled receptor for oleoylethanolamide and its use in the discovery of small-molecule hypophagic agents. *Cell. Metab.* **2006**, *3*, 167–175.

(77) Felder, C. C.; Briley, E. M.; Axelrod, J.; Simpson, J. T.; Mackie, K.; Devane, W. A. Anandamide, an endogenous cannabimimetic eicosanoid, binds to the cloned human cannabinoid receptor and stimulates receptor-mediated signal transduction. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 7656–7660.

(78) Irwin, J. J. Community benchmarks for virtual screening. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 193–199.

(79) He, W.; Miao, F. J.; Lin, D. C.; Schwandner, R. T.; Wang, Z.; Gao, J.; Chen, J. L.; Tian, H.; Ling, L. Citric acid cycle intermediates as ligands for orphan G-protein-coupled receptors. *Nature* **2004**, *429*, 188–193.

(80) Brown, A. J.; Goldsworthy, S. M.; Barnes, A. A.; Eilert, M. M.; Tcheang, L.; Daniels, D.; Muir, A. I.; Wigglesworth, M. J.; Kinghorn, I.; Fraser, N. J.; Pike, N. B.; Strum, J. C.; Steplewski, K. M.; Murdock, P. R.; Holder, J. C.; Marshall, F. H.; Szekeres, P. G.; Wilson, S.; Ignar, D. M.; Foord, S. M.; Wise, A.; Dowell, S. J. The Orphan G protein-coupled receptors GPR41 and GPR43 are activated by propionate and other short chain carboxylic acids. *J. Biol. Chem.* **2003**, *278*, 11312–11319.

(81) Barillari, C.; Marcou, G.; Rognan, D. 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. *J. Chem. Inf. Model.* **2008**, *48*, 1396–1410.

(82) Hoberman, R.; Klein-Seetharaman, J.; Rosenfeld, R. Inferring property selection pressure from positional residue conservation. *Appl. Bioinf.* **2004**, *3*, 167–179.

(83) Vert, J. P.; Jacob, L. Machine learning for in silico virtual screening and chemical genomics: new strategies. *Comb. Chem. High Throughput Screening* **2008**, *11*, 677–85.

(84) Bingham, J.; Sudarsanam, S. Visualizing large hierarchical clusters in hyperbolic space. *Bioinformatics* **2000**, *16*, 660–661.

CI800447G