JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

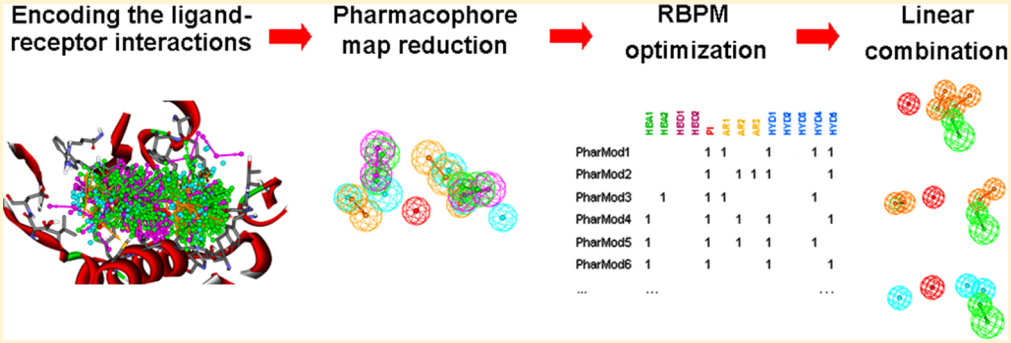
pubs.acs.org/jcim

New Strategy for Receptor-Based Pharmacophore Query Construction: A Case Study for 5-HT$_7$ Receptor Ligands

Rafał Kurczab and Andrzej J. Bojarski*

Department of Medicinal Chemistry, Institute of Pharmacology Polish Academy of Sciences, 12 Smętna Street, Kraków, 31-343, Poland

Supporting Information



ABSTRACT: In this paper, a new approach for generating receptor-based 3D pharmacophore models for rapid *in silico* virtual screening is presented. The method combines information from docking poses of known ligands of different structures and further ligand−receptor complexes analyses using structural interaction fingerprints (SIFts). Next, the best linear combination of three-, four-, and five-feature pharmacophores in terms of selected performance parameter (i.e., recall, F-score, and MCC) is constructed. The resultant queries showed significantly better VS performance and new scaffold recognition when compared with the known ligand- and receptor-based pharmacophore models. The approach was developed and validated on 5-HT$_7$ receptor homology models created on available crystal structure templates. The efficiency of the obtained linear combinations exhibited only a minor dependence on the template selection.

INTRODUCTION

Pharmacophore models have been successfully used in lead generation, scaffold hopping, mining small molecule databases, and other drug design applications.[1,2] The pharmacophore describes the essential steric and electronic function-determining points necessary for optimal ligand interactions with a given receptor, thereby eliciting similar chemical and biological activity.[3] There are several methods for generating pharmacophores, depending on the available information. When a number of ligands of a given target is known, the ligand-based pharmacophore models can be created. This approach is of particular relevance for targets that are difficult to crystallize, such as membrane proteins (e.g., ion channels, transporters, or G protein-coupled receptors (GPCRs)). When only the structure of a receptor is available (e.g., orphan receptors), structure-based pharmacophore generation methods, which identify possible interaction sites within a binding site, are used. If the structure of a receptor and its ligands are known, both of the above-mentioned methods can be used, or a hybrid pharmacophore that involves information from both sources can be constructed. This approach is also called receptor- or contact-based pharmacophore generation.

Because our new pharmacophore generation method is based on the structure of a target protein, other similar approaches reported thus far are briefly reviewed below.

Regarding the construction of structure-based pharmacophore models in which no prior knowledge of ligands or their interactions is used, Sanders et al. presented a tool for class A GPCRs called Snooker.[4] The pharmacophore hypotheses were obtained from homology models created on-the-fly based on the sequence alignment of the seven transmembrane domains and a crystal structure template. Residues important for ligand binding were identified by analysis of Shannon entropies of structurally conserved positions in a multiple sequence alignment, and the chemical features representing protein−ligand interactions were positioned in the binding pocket.

Another method of this type has been published by Klabunde et al.,[5] who showed an automated preparation of sequence-derived 3D-structure-based pharmacophore hypothesis based on chemoprints originating from 10 different homology models and three X-ray structures of class A GPCRs.

One of the first approaches to structure-based pharmacophore model generation was the Pocket module of LigBuilder,[6]

Received: September 6, 2013
Published: November 18, 2013

dx.doi.org/10.1021/ci4005207 | J. Chem. Inf. Model. 2013, 53, 3233−3243

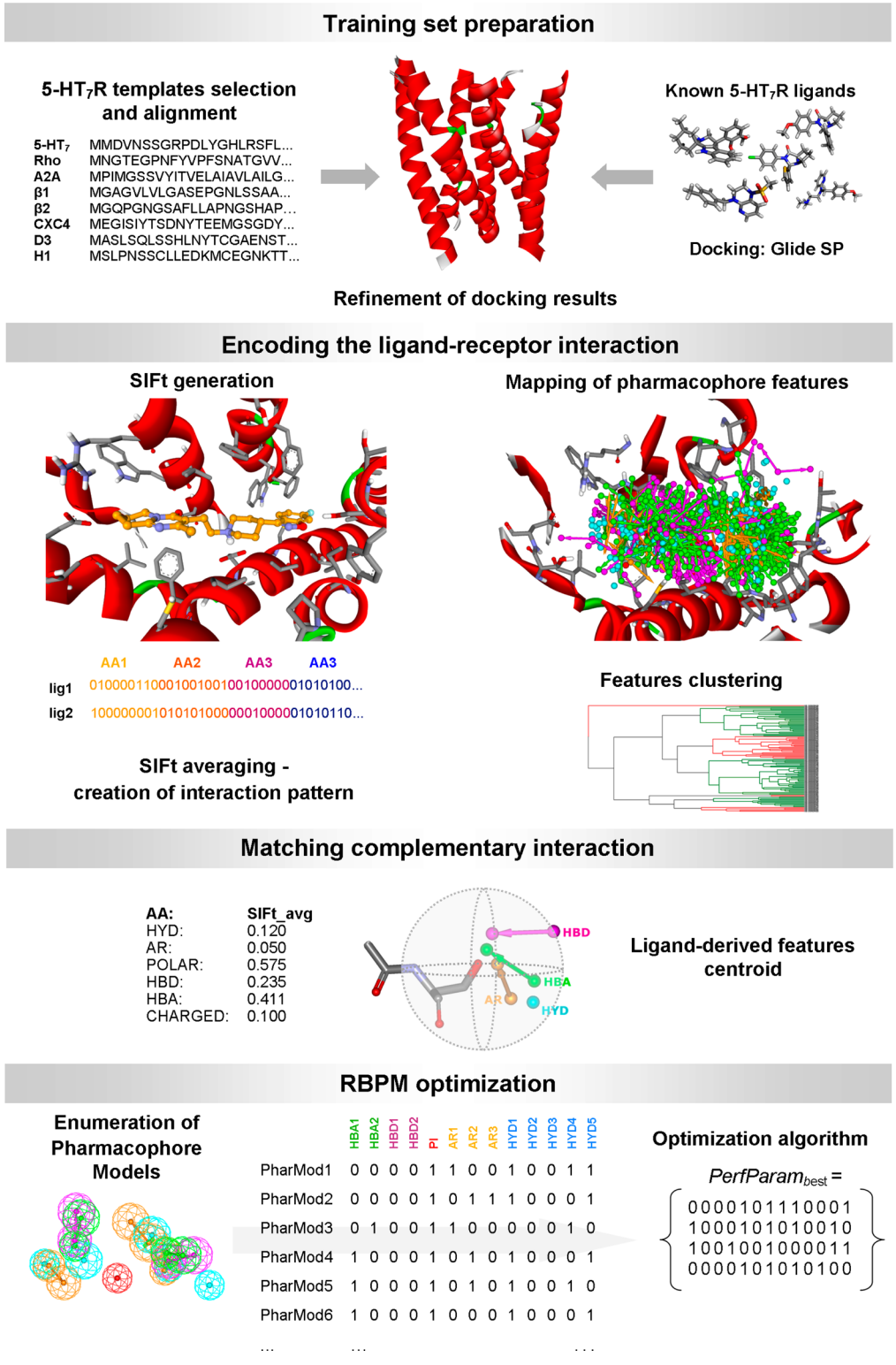


Figure 1. Flowchart describing the developed method of receptor-based pharmacophore generation.

a multipurpose program for *de novo* structure-based ligand design used to analyze the binding site by means of scored grids and extract a receptor-based pharmacophore model. Due to difficulties with the selection of crucial features among a number of centers, an extension denoted Pocket v.2 was released.[7] This extension used the ligand−receptor complex to mark all contributing spots, which were then used to obtain the binding intensity, assuming that the decisive pharmacophore features are responsible for affinity.

Wolber and Langer developed the LigandScout algorithm, which acquired information from ligand-protein complexes stored in the PDB archive.[8] They defined a set of rules that automatically detected and classified the intermolecular contacts into hydrogen bond interactions, charge transfers,

dx.doi.org/10.1021/ci4005207 | J. Chem. Inf. Model. 2013, 53, 3233−3243

and lipophilic regions. The entire set of interactions formed a pharmacophore model that can be used in virtual screening.

Starting from the Protein Data Bank (PDB) macromolecule complex and using the GRID approach, Ortuso et al.[9] developed a general and automatic computational procedure that generated the so-called GRID-based pharmacophore model (GBPM). They logically combined GRID force field maps[10] to derive necessary information on the interactions from a ligand−receptor complex. The preliminary models were converted from the GBPM points into the Catalyst features and were evaluated in the CiTest fit module.

A different method, known as HS-Pharm,[11] created pharmacophore models from a machine learning algorithm trained on known cavity fingerprints. The identification of the most interacting ligand atoms was followed by construction of receptor-based pharmacophores from the atom-probe interaction maps. Next, Meslamani et al.[12] described a fully automated method for the construction of 3D pharmacophore queries from ligand-protein structures. This Receptor−Ligand Pharmacophore Generation (RLPG) protocol, implemented in Discovery Studio 3.1,[13] used features that corresponded with the ligand−receptor interactions and enumerated candidate pharmacophore models. The hypotheses were then ranked by their selectivity predicted by a Genetic Function Approximation (GFA) model.[14]

Another protocol was introduced by Salam et al.,[15] who obtained energy-optimized pharmacophores (e-pharmacophores) based on mapping of the energetic terms from the Glide XP scoring function onto atom centers. Next, pharmacophore sites were composed, and the Glide XP energies from the atoms were summed. The pharmacophore features were then ranked, and the most energetically favorable features were selected for the final pharmacophore hypothesis.

All of the above-mentioned receptor-based pharmacophore generation methods are run as single-batch mode methods; i.e., these methods are able to produce and select the best pharmacophore models based only on a single L-R complex, leading to hypotheses limited only to one chemical class of ligands. Certainly, for each known chemotype, a separate model can be prepared and used, e.g., in virtual screening; although such an approach may be efficient for limited ligands diversity, otherwise it is too expensive computationally. Therefore, we developed a new approach to the automated generation of 3D receptor-based pharmacophore models by using a diverse set of actives to probe the binding site of a receptor. This approach produced a comprehensive pharmacophore features map from which a linear combination of only several (3−8) pharmacophore models is able to match different chemical scaffolds of actives and return a low number of false positives. The approach was developed on a serotonergic 5-HT_7 receptor, a member of class A GCPRs, and a diversified set of 5-HT_7R ligands. Due to the availability of many different crystal structures of GPCRs, the influence of receptor template selection on the quality and virtual screening efficacy of the produced pharmacophore models was also investigated.

■ MATERIALS AND METHODS

The proposed methodology consisted of several stages schematically represented in Figure 1. The basic idea is to use information from important pharmacophore points at a receptor and ligand level. To achieve this goal, the structural interaction fingerprint (SIFt) profiles and ligand-based pharmacophore maps were created and analyzed to match complementary interactions. Those filtered features created a set of important anchoring interaction points, of which single pharmacophore models and then a linear combination of pharmacophore models are constructed so as to maximize the parameters of search performance.

Training and Test Sets Preparation. Structures of the 5-HT_7R ligands were extracted from ChEMBL v10 database[16] using an activity threshold K_i less than 300 nM. The obtained set of 230 compounds was next hierarchically clustered using Molprint2D fingerprint and Tanimoto metric implemented in Canvas,[17] producing 28 groups. Of the 19 clusters, the centroid and a few (proportional to a cluster size) of the most structurally diverse agents formed the training set of 55 ligands, whereas the rest were included in the test set. The decoy sets were created by following the DUD methodology.[18] They were selected from ZINC database[19] and span the same structural property ranges as known actives (molecular weight, H-bond donor count, H-bond acceptor count, logP, and basic pK_a). Next, to remove decoys structurally close to any of true actives, a maximal Tanimoto similarity threshold of 0.3, calculated for a Daylight-type chemical hashed fingerprint, was applied. The ratio between actives and decoys was set to 1:10.

For each compound from a training and test sets, a maximum of 250 conformers were generated in the Discovery Studio 2.5 Catalyst module using the FAST[20] parameter settings, with an upper energy threshold of 20 kcal mol^{-1}.

Generation of 5-HT_7R Homology Models. To receive reliable models of the target, a series of homology models was prepared. The sequence of the human 5-HT_7 (ID: P34969) receptor was obtained from the UniProtKB/Swiss-Prot database.[21] The crystal structures of GPCRs were downloaded from the PDB repository: adenosine 2 receptor (PDB ID: 3QAK), β1 (PDB ID: 2Y00) and β2 (PDB ID: 3P0G) adrenergic receptors, CXC chemokine receptor type 4 (PDB ID: 3OE0), dopamine 3 receptor (PDB ID: 3PBL), histamine 1 receptor (PDB ID: 3RZE), and rhodopsin (PDB ID: 1F88). Sequence alignments were performed using the Discovery Studio 2.5 package. The prediction of ranges of helices was supported by Web services: PONGO,[22] ExPASY PROSITE,[23] and psipred.[24] For simplicity, the extracellular loops were omitted, and a series of 200 models was generated by MODELLER 9v8 software for each aligned template.[25]

Selection of Homology Models. The obtained models were next evaluated to select the most suitable conformations for pharmacophore preparation. The Protein Preparation Wizard was used to assign the bond orders, check the steric clashes, and assign appropriate amino acid ionization states. The receptor grids were generated (the OPLS_2005 force field) using the same size and position of the grid box centered on the Asp3.32. In the first stage, a set of structurally diverse 5-HT_7R active ligands[26] were docked using Glide XP mode.[27] Receptor models accommodating less than half of the ligands with average score less than −3.0 were rejected from further evaluation. In the second step, an extended set of ligands was docked using Glide with standard precision mode (SP). This set, apart from structurally similar ligands to those from the initial set, contained additional ones having different chemical scaffolds. The two best receptors for each template were picked for further research from the models accommodating most of the ligands with a Glide score better than −6.0.

Docking and Pose Refinement. The cross-docking of the training set to a collection of the selected homology models was performed using the Glide docking algorithm on SP mode. The

dx.doi.org/10.1021/ci4005207 | J. Chem. Inf. Model. 2013, 53, 3233−3243

ligand ionization states at pH = 7.4 were assigned using Epik.[28] The docking poses interacting with Asp3.32 different than by a salt bridge were removed (e.g., amide, hydroxyl group, and carbonyl group).

Generation and Analysis of SIFt-Based Interaction Patterns. The structural interaction fingerprint method was used to identify amino acids that interact with the complexed ligand. SIFts were calculated using our in-house implementation of the algorithm based on methodology proposed by Deng et al.[29,30] The results were stored in a 1D binary string, in which a 9-bit pattern was used to describe the interaction type: any contact, backbone, side chain, polar, aromatic, hydrophobic interaction, hydrogen bond donor/acceptor, and charged. Then, for each receptor model, the normalized frequencies of the given interaction were calculated, and amino acids involved in less than 10% of complexes were completely removed from the further study. The two sets of interaction patterns were considered, one containing all interacting amino acids (max mode) and a second focused only on those for which interaction frequencies were higher than 50% (min mode).

Pharmacophore Feature Mapping and Clustering. The docked ligand conformations were mapped to a set of pharmacophore features, namely hydrogen bond acceptor (HBA) and donor (HBD), positive ionizable group (PI), hydrophobic region (HYD), and aromatic ring (AR), using the Feature Mapping tool of Discovery Studio 2.5. Within the obtained comprehensive map of spatial distribution of various pharmacophore points in the binding site, the same type of features were then clustered taking into account distances between all possible pairs of feature centroids. Next, the densest clusters (containing at least five features) were selected, and their average location was calculated.

Matching Complementary Interaction. The data derived from SIFt (from the receptor side) and Catalyst's feature distributions (from the ligands side) were next concatenated by application of an automatic algorithm detecting and recognizing the most important molecular interactions. The set of key rules described by Wolber and Langer[8] and one extra rule, which defined dipole−dipole contacts, were applied (more in Appendix 1, Supporting Information). For amino acids identified by SIFt pattern analysis, a sphere of 6 Å radius was then scanned, and the features that did not match the interaction rules were removed. The simplified feature-to-feature (without geometry constrains) recognition rules used at this step are summarized in Table 1.

Enumeration of Pharmacophore Models. The Screening Library protocol from Discovery Studio 2.5 was used to enumerate a subset of possible pharmacophores from a final comprehensive features map using an active subset of the training set. For each case, the maximum subset of hypotheses was set to 150, and only three-, four-, and five-feature pharmacophore hypotheses were generated (see Appendix 2, Supporting Information). The options, minimum interfeature distance of 0.5 Å and rigid fitting, were additionally enabled.

Table 1. Matching Complementary Interaction Rules Used To Reduce the Number of Features

interaction indicated from		
receptor view (SIFt)	ligand view (Catalyst)	final feature
hydrophobic (HYD)	hydrophobic (HYD)	hydrophobic (HYD)
aromatic (AR)	aromatic (AR)	aromatic (AR)
H-bond acceptor (HBA)	H-bond donor (HBD)	H-bond donor (HBD)
H-bond donor (HBD)	H-bond acceptor (HBA)	H-bond acceptor (HBA)
charged	positive ionizable (PI)	positive ionizable (PI)
	negative ionizable (NI)	negative ionizable (NI)
polar	{HBA, HBD}	{HBD, HBA}

Assessment of Method Performance. For the evaluation of the obtained models, a set of measures was used: recall (R), precision (P), specificity (Sp), enrichment factor (EF), F-score (Fs), and the Matthews Correlation Coefficient (MCC):[31]

$$R = \frac{TP}{TP + FN} \quad (1)$$

$$P = \frac{TP}{TP + FP} \quad (2)$$

$$Sp = \frac{TN}{TN + FP} \quad (3)$$

$$EF = \frac{TP(TP + FP + TN + FN)}{(TP + FP)(TP + FN)} \quad (4)$$

$$Fs = \frac{2PR}{P + R} = \frac{2TP}{2TP + FP + FN} \quad (5)$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \quad (6)$$

Recall (1) measures the amount of positive elements correctly classified, precision (2) describes the correctness of positive prediction, and specificity (3) measures the proportion of negatives that are correctly identified. The F-score (or F-measure, 5) is a measure of a test's accuracy, which varies from 1 to 0. The F-score considers both the precision and recall of the test and can be interpreted as a weighted average of these values. The MCC (6) is a balanced measure of binary classification effectiveness ranging from −1 to 1, where 1 represents perfect prediction.

The precision-recall plot[32] is often used in information retrieval to assess the algorithm in binary classification test. The chart area can be divided into four parts, grouping the models according to their two-dimensional performance.

Optimization of RBPM Performance. An output sdf file from Screening Library and the in-house scripts were used to extract data for further analysis (i.e., list of unique PharmPrints,[33,34] the definition of bit positions in PhramPrint, and the ID of compounds matching on a given pharmacophore). This was supported by the MayaChemTools (Perl)[35] and JChem (Java)[36] libraries. Next, the obtained data were converted to a binary matrix (see Appendix 3, Supporting Information) encoding mapping of the training compounds on the given model. This matrix served as the input to the algorithm searching the minimal subset of pharmacophore models that produced the highest improvement in performance. The best minimal combinations of pharmacophore models found by the optimization algorithm were validated on the test set. All of the calculations were performed on an Intel Core i7 CPU 3.00 GHz computer system with 24 GB RAM running a 64-bit Linux operating system.

Reproduction of the Known 5-HT$_7$R Pharmacophore Models. The known ligand-based 5-HT$_7$R pharmacophore models (see Appendix 4, Supporting Information) were reconstructed based on the described methodology[37,38] using the Catalyst module of Discovery Studio 2.5. The obtained

dx.doi.org/10.1021/ci4005207 | J. Chem. Inf. Model. 2013, 53, 3233−3243

models satisfactorily fit the geometrical (i.e., interfeature distances) and pharmacophore (i.e., count and type of features) definitions.

From the general hypothesis of the receptor-based pharmacophore model proposed by Kołaczkowski et al.,[26] the eight different variations of features were retrieved and stored as maps (Appendix 4, Supporting Information) in Discovery Studio. Each map was composed with eight pharmacophore points corresponding to the observed ligand−receptor interactions. To assess the performance of a model, the simplified scheme was used, i.e., the enumeration of three- to five-feature pharmacophores and searching for the best linear combination. Two extreme models (the best, ref_RB_best, and the worst, ref_RB_worst) were taken as a representative subset of receptor-based models for comparative studies.

RESULTS AND DISCUSSION

The 5-HT$_7$ receptor-based pharmacophore models obtained with the proposed methodology were analyzed in relation to known hypotheses, and because these models were built on different templates, their influence on model performance was evaluated. Two approximation modes were considered, one based on all identified L-R interactions (max mode) and the second focusing on highly important contacts (min mode). Next, due to a large number of possible pharmacophore models produced for a single receptor conformation, an in-house algorithm searching of their minimal combination maximizing different classification parameters was developed. At the final stage, nine clusters of actives that were not used in model creation were used to evaluate both single models and the best combinations in discovering new chemotypes.

Enumeration of the Pharmacophore Models Subset. A feature map was obtained for each receptor model and approximation mode, and the number of features ranged from 8−15 and 13−17 for the min and max modes, respectively. Switching on the min mode caused a reduction of final maps from 0 (for β1_110, the number of features in min and max mode was the same) to 7 pharmacophore points, and no significant tendency in pharmacophore type's distribution was observed. Of each map, an initial subset of 150 pharmacophore models was created by directly linking features in the three- to five-point hypotheses with the Screen Library protocol from Discovery Studio 2.5. The population of three-point models ranged from 5.1% and 6.0% to 19.8% and 31.7% for maps with max (D3_017_max and CXC4_028_max) and min (β2_067_min and D3_107_min) number of features, respectively (Appendix 2, Supporting Information). Four-feature models formed a stable fraction (31%−41%), whereas the amount of five-feature models (28%−64%) increased with the size of the feature map.

Because the set of compounds used in sampling of the binding site was constant, the receptor conformation had a predominant influence on feature map composition. The analysis of SIFt interaction patterns for actives and inactives separately showed that some types of L-R contacts were preferred, regardless of the type of template and receptor conformation selection. Using this data, the differential distribution of features between actives and decoys was calculated. Figure 2 presents an example diagram obtained for H1_123_min, where four features were more frequently matched by actives than by decoys.

A similar methodology was used in HS-Pharm development,[11] in which such diagrams helped constraining features in the final hypothesis. Here, this concept was employed to reduce the number of considered pharmacophores to save time during the optimization of linear combination and produce more selective models. Detailed analysis of all the obtained diagrams indicated the two most selective features: PI and one AR. They correspond with the well-recognized key interactions for 5-HT$_7$R (as well as other aminergic GPCRs), i.e., creation of a salt bridge between the charged moiety of a ligand (PI) and Asp3.32 and $\pi-\pi$ aromatic interaction (AR) with Phe6.51 and/or Phe6.52.[26] Because Asp3.32 is considered to be the main anchoring point for a ligand and has the highest impact on active/decoy separation, it was used as a constraint. These results were also in line with known ligand-based pharmacophore models, in which the PI feature was reported to be crucial for high sensitivity of the 5-HT$_7$R models.[37−39]

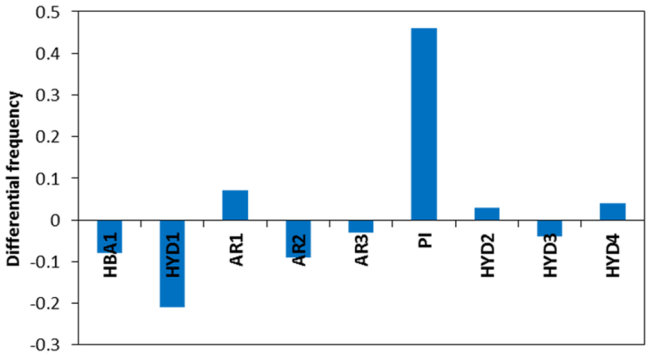


Figure 2. Diagram showing the differential distribution of particular feature between actives and decoys.

The Best Single Model. Using the feature constraint, all of the initial subsets of 150 models were reduced to those containing positive ionizable pharmacophore point. Then, for each pharmacophore model, different parameters evaluating performance were calculated. Table 2 shows a juxtaposition of hypotheses with the highest MCC values compared with the reference ligand- and receptor-based pharmacophore models.

The global performance parameters (MCC and Fs) had relatively narrow ranges: 0.44−0.62 and 0.40−0.59, respectively, and they both indicated nearly the same models from a given template as top- and bottom-ranked. The precision-recall plot (Figure 3) showed that high MCC values resulted from either (i) high recall but low precision (e.g., β2_067_max), (ii) high precision but low recall (e.g., CXC4_028_max), or (iii) both recall and precision >0.5 (eight models: three on H1, two on D3 and β1, and one on the β2 template). It should be stressed that despite the template used, all new hypotheses (except β1_115_max) outperformed reference ligand- and even receptor-based models. Although only one of the new pharmacophores (D3_107_min) was composed of the same features as two reference models, it showed higher values for each parameter, confirming that the proposed approach is able to optimize the mutual orientation between features more accurately.

Regarding the number of features, the total amount of three-, four-, and five-feature models was 3, 14, and 10, respectively. This is in line with another study[11] that showed that screening with four-feature pharmacophores is the best compromise between three-feature pharmacophores, which are usually too permissive (high recall and low precision), and screening with five-feature pharmacophores, which are often too restrictive (low recall and high precision). It is also worth noting that the

dx.doi.org/10.1021/ci4005207 | J. Chem. Inf. Model. 2013, 53, 3233−3243

Table 2. Results of Virtual Screening Performance for Pharmacophore Models with the Highest MCC Value (Selected for a Given Receptor Template and Conformation) and the Reference Pharmacophores

model ID	R	P	Sp	EF	Fs	MCC	hypothesis[a]
A2A_006_max	0.69	0.44	0.90	4.44	0.53	0.49	PHHH
A2A_006_min	0.42	0.70	0.98	7.14	0.53	0.51	PRHHH
A2A_014_max	0.82	0.44	0.88	4.41	0.57	0.54	PRHH
A2A_014_min	0.83	0.43	0.88	4.38	0.57	0.54	PRHH
β1_110[b]	0.54	0.58	0.96	5.87	0.56	0.51	PRRHH
β1_115_max	0.39	0.52	0.96	5.22	0.44	0.40	PRHHH
β1_115_min	0.54	0.66	0.97	6.66	0.59	0.55	PRRH
β2_034_max	0.86	0.42	0.87	4.31	0.57	0.55	PHH
β2_034_min	0.77	0.45	0.90	4.56	0.57	0.53	PRHH
β2_067_max	0.88	0.44	0.88	4.45	0.59	0.57	PRHH
β2 067_min	0.82	0.45	0.89	4.60	0.59	0.56	PRHH
CXC4_028_max	0.31	0.83	0.99	8.45	0.46	0.48	PRHH
CXC4_028_min	0.49	0.67	0.97	6.81	0.57	0.54	APRHH
CXC4_152_max	0.54	0.65	0.97	6.55	0.59	0.55	PRHH
CXC4_152_min	0.77	0.47	0.91	4.80	0.59	0.55	PHH
D3_017_max	0.42	0.71	0.98	7.21	0.53	0.51	PRHH
D3_017_min	0.57	0.68	0.97	6.87	0.62	0.58	PRRH
D3_107_max	0.83	0.43	0.88	4.38	0.57	0.53	AAPHH
D3_107_min	0.57	0.52	0.94	5.28	0.55	0.49	APRH
H1_083_max	0.81	0.37	0.85	3.74	0.51	0.48	APH
H1_083_min	0.53	0.73	0.98	7.10	0.61	0.59	PRRHH
H1_123_max	0.52	0.69	0.98	7.04	0.59	0.56	PRRHH
H1_123_min	0.5	0.75	0.98	7.60	0.60	0.58	PRRHH
Rho_175_max	0.49	0.67	0.97	6.76	0.57	0.53	APRRH
Rho_175_min	0.42	0.64	0.97	6.49	0.51	0.48	APRRH
Rho_149_max	0.76	0.46	0.90	4.70	0.58	0.53	PRHH
Rho_149_min	0.59	0.49	0.93	4.95	0.54	0.48	DPHH
ref_RB_best[c]	0.47	0.49	0.95	4.95	0.48	0.43	APRH
ref_RB_worst[d]	0.11	0.83	0.99	8.45	0.20	0.29	APHH
ref_LB_2000[e]	0.54	0.45	0.93	4.60	0.49	0.43	APRH
ref_LB_2003 [f]	0.35	0.34	0.93	3.49	0.35	0.27	APHHH

[a]Letter abbreviations used for coding features in pharmacophore hypotheses: A – H-bond acceptor, D – H-bond donor, P – positive ionizable group, R – aromatic ring, H – hydrophobic region. [b]For the β1_110 template, the composition of pharmacophore maps in min and max mode was the same. [c]The best single hypotheses obtained on two selected pharmacophore maps retrieved from Kołaczkowski et al.[26] [d]The worst single hypotheses obtained on two selected pharmacophore maps retrieved from Kołaczkowski et al.[26] [e]Ligand-based pharmacophore models reconstructed from ref 37. [f]Ligand-based pharmacophore models reconstructed from ref 38.

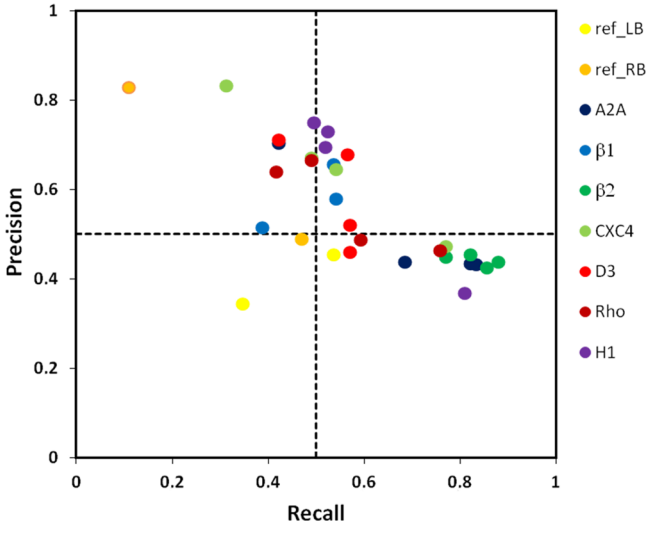


Figure 3. The precision-recall plot for the best single hypotheses.

most frequently occurring features composition, PRHH, was found nine times and resulted in models with the best recall (β2_067_max), precision, specificity, and enrichment factor (CXC4_028_max).

Searching for the Best Linear Combination. Because a single hypothesis is usually unable to fully cover the chemical space of ligand structures, successive efforts were directed to create a comprehensive pharmacophore query composed of the minimal number of single models needed to ensure the best possible screening performance. Linear combinations with the highest MCC values composed of the proposed algorithm (see Appendix 3, Supporting Information) are summarized in Table 3. As expected, linear combinations provided better results than any single model ($0.60 < MCC < 0.66$); thus, they outperformed the reference ligand-based models and were also superior to any combination of pharmacophore features defined by Kołaczkowski et al.[26] Interestingly, similar to single hypotheses, the best combinations were obtained on the H1 and β2 templates and used a reduced set of interactions (min mode). Detailed inspection of optimization results showed that in many cases, the best single hypothesis was not a part of the best linear combination.

In the precision-recall plot (Figure 4), nearly all combinations occupied the region of the best classification ability, and

dx.doi.org/10.1021/ci4005207 | J. Chem. Inf. Model. 2013, 53, 3233–3243

Table 3. Best Linear Combinations for a Given Template Selected among Results Obtained for Different Conformations and Two Modes

model ID	option	R	P	Sp	EF	Fs	MCC	size
Rho_149_min	MCC[a]	0.74	0.62	0.95	6.31	0.68	0.64	6
A2A_006_min	MCC	0.86	0.49	0.90	4.99	0.63	0.60	6
D3_017_min	MCC[a]	0.61	0.68	0.97	6.84	0.64	0.60	4
CXC4_028_min	MCC[a]	0.79	0.56	0.93	5.64	0.65	0.62	4
β1_110	MCC[a]	0.81	0.57	0.93	5.81	0.67	0.64	5
β2_067_max	MCC[a]	0.68	0.69	0.97	7.01	0.69	0.65	6
H1_123_min	MCC	0.72	0.66	0.96	6.72	0.69	0.66	5
ref_RB_best	MCC	0.71	0.56	0.92	5.68	0.63	0.59	7
ref_RB_worst	MCC	0.44	0.57	0.96	5.83	0.50	0.46	10

[a]In this case, the same combination was also found by optimizing the F-score value.

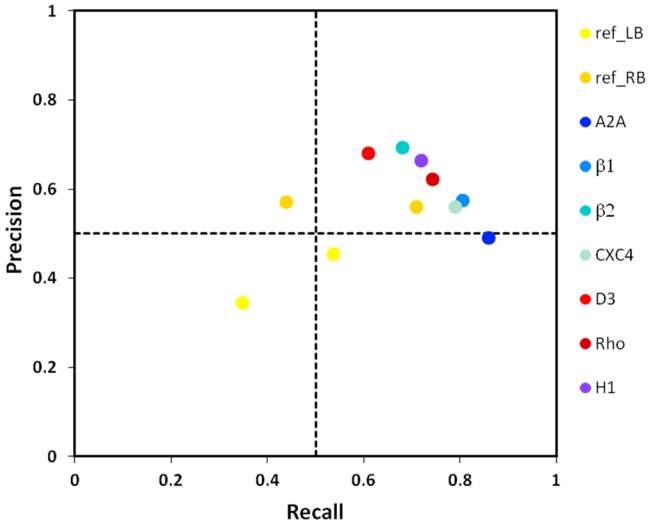


Figure 4. Plot illustrating the performance of the best linear combinations in precision-recall space.

the differences between them were relatively small because the final query, which included several different models, can reach similar level of activity pattern recognition. The classic ligand-based approach is not capable of creating such a broad (though limited) set of pharmacophore models in a batch mode. Even using diverse set of ligands, there is a problem of selecting models to be used to screen a library. One can generate pharmacophore models for each structural class and then use them to search for the best linear combination to achieve the similar effect. Because this process is not automated, this solution can be too time-consuming for a large number of actives, primarily due to the necessity of generating a large number of conformations and further building and scoring the pharmacophore models.

A closer examination of the best combination (Figure 5) showed that it was composed of different pharmacophore subsets capable of covering a wide variety of chemical scaffolds.

The model presented in Figure 5A contains three features and is an example of the general hypothesis, which usually exhibits high recall but low specificity. The second model (Figure 5C) is characterized by almost overlapped aromatic ring and hydrophobic feature (average distance = ~0.6 Å) that expands the searching area during ligands mapping. The definitions of hydrophobic and aromatic ring features in Discovery Studio 2.5 allowed the detection of the aromatic moiety by both features; however, hydrophobic rules can find appropriate substituent connected to the ring, which together provide a very sensitive and specific model. Interestingly, a similar configuration was found in several of the final combinations, as well as in the best single models. Two hypotheses (Figure 5D and 5E) had very similar topology to the ligand-based models proposed by Lopez-Rodriguez, with an aromatic instead of hydrophobic feature and the location of one hydrophobic area for 5D vs ref_LB_2000 and 5E vs ref_LB_2003, respectively. The last model lacks the hydrophobic or ring aromatic features located at the left-hand side of positive ionizable group, making it perfect to match short and bulky 5-HT7R ligands, such as some ergolines, aporphines, tricyclic psychotrope derivatives, etc.

Max vs Min Mode. Regarding the influence of feature map reduction on the performance of final combinations, the differences between min and max mode for all evaluative measures were calculated. The most interesting fluctuations were illustrated in Figure 6, in which the MCC and F-score were used to monitor the global behavior of models. To compare the three optimization strategies used, the blue bars represent the changes of particular parameter between combinations found by MCC maximization, and the red and green bars are based on F-score and recall, respectively. The reduced set of interaction points led to the enhancement of the linear combinations of pharmacophore models. In general, the increase of performance varied from 1% to 10% and strongly depended on the optimized parameter. The linear combinations optimized for full and reduced feature maps using recall produced the lowest increase in performance. Moreover, recall optimization made the worst models in two cases (D3_017, β2_067), and, in one case, no improvement was observed. The results for MCC and F-score were rather consistent, and their optimization had the highest impact on the enhanced performance of the final combinations.

Recognition of New Chemical Scaffolds. The nine clusters of 5-HT$_7$R actives not used in models generation were used to evaluate the proposed approach in light of their ability to discover new chemical scaffolds. The cluster centroids are presented in Figure 7 together with their affinity and database identification number. There were several singleton clusters (i.e., 10, 16, and 17), and the others were represented by a few derivatives (2–4). If at least one member was identified by the hypothesis, then it (chemotype) was considered recognized. Table 4 presents the number of clusters identified by the best single query, the linear combinations optimized on the given parameter, and the reference models.

In general, the best single hypotheses as well as the ligand-based models showed poor ability to identify the unknown active structures; however, some models (i.e., CXC4_152_min,

dx.doi.org/10.1021/ci4005207 | J. Chem. Inf. Model. 2013, 53, 3233−3243

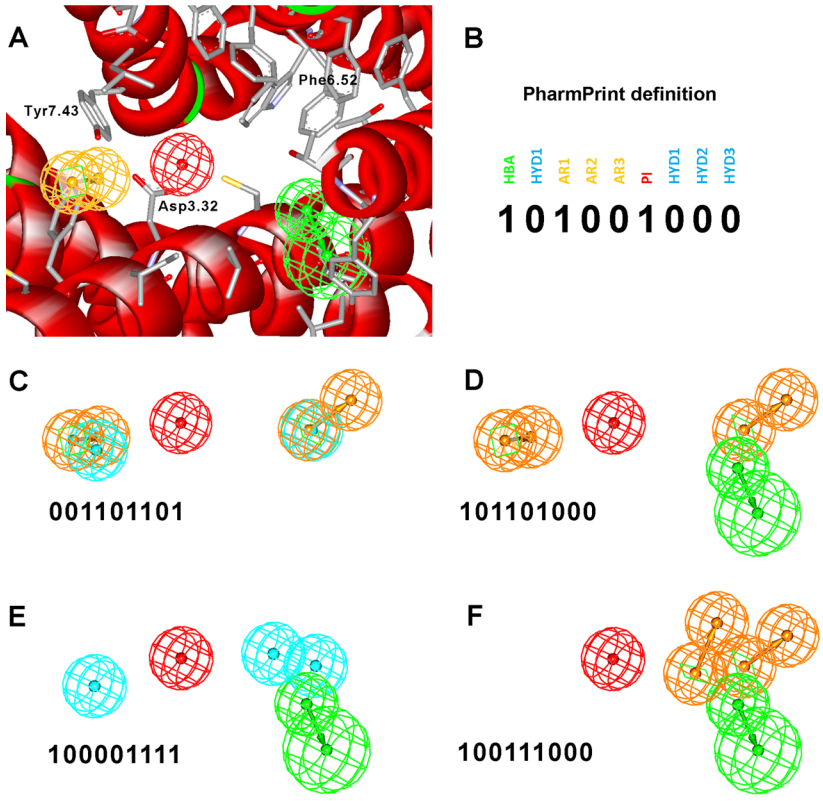


Figure 5. Graphical representation of the best linear combination for the H1_123_min model.

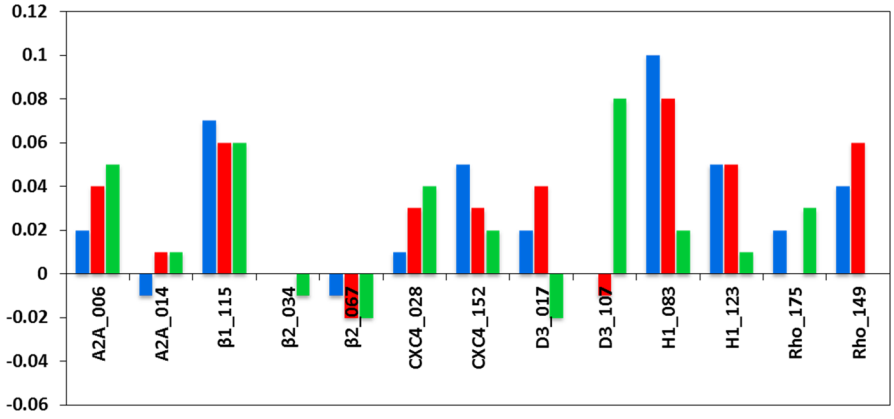


Figure 6. Differential plot calculated for pharmacophore linear combinations built on the min and max modes. The color of the bars indicates the type of performance parameter used to find the best combination (blue − MCC, red − F-score, and green − recall).

H1_083_max) recovered a full set of scaffolds. In the best single query, the max mode showed a better general performance compared with the min approach.

The linear combination enhanced the performance of pharmacophore search of the database. In our tests, the strategy based on recall optimization showed the best effectiveness in the recognition of new active scaffolds, and, only in two cases, this level was slightly lower than 100%. Unfortunately, the combinations optimized by the use of MCC or the F-score had a slightly worse performance compared with recall. In three cases (H1_083_max and *β*2_067_min, *β*2_067_max) they were even less effective than the best single hypothesis for a particular receptor model. These combinations, however, did not contain the best single model. Interestingly, in those cases, several combinations with equally high MCCs were found, and some of them showed a better new scaffold recognition level (however, they were characterized by lower values in other performance parameters). This confirms that the applied optimization method can be further improved, for example by implementation of multivariate optimization techniques.

CONCLUSION

The primary objective of this study was to develop and verify a new approach to generate the extended pharmacophore query, enabling fast and comprehensive mining of the chemical space in search of structurally new active compounds. The proposed method is based on the information obtained from probing of the binding site by known ligands and analysis of L-R interaction patterns obtained using the SIFt approach. These

dx.doi.org/10.1021/ci4005207 | J. Chem. Inf. Model. 2013, 53, 3233−3243

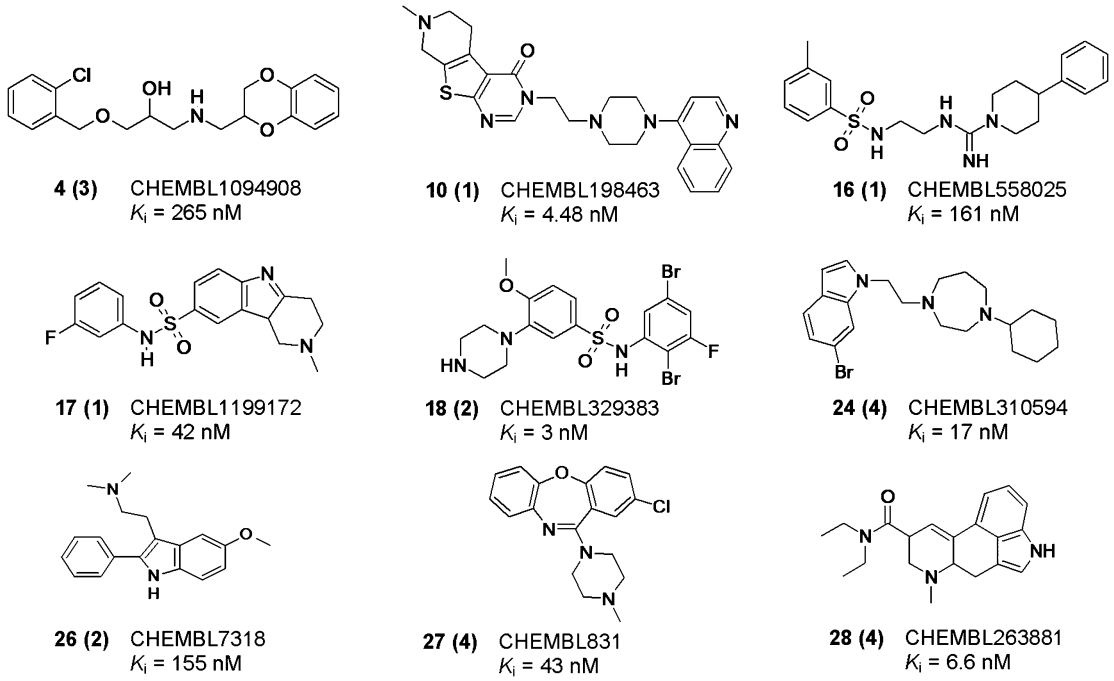


Figure 7. The centroids of clusters used in the scaffold recognition test (the number of cluster members was added into parentheses).

Table 4. Effectiveness of New Active Scaffold Recognition in the Virtual Screening Test

	best single			linear combination		
model ID	R	Fs	MCC	R	Fs	MCC
A2A_006_max	6	6	6	9	8	8
A2A_006_min	4	3	3	9	8	7
A2A_014_max	7	7	7	9	9	9
A2A_014_min	7	7	7	9	7	9
β1_110	7	5	5	9	6	6
β1_115_max	5	5	5	7	8	8
β1_115_min	8	3	3	9	8	9
β2_034_max	8	8	8	9	9	9
β2_034_min	7	7	7	9	8	9
β2_067_max	8	8	8	9	6	6
β2_067_min	8	8	8	9	6	6
CXC4_028_max	7	7	1	9	5	5
CXC4_028_min	5	3	3	9	8	8
CXC4_152_max	8	3	3	9	5	9
CXC4_152_min	9	9	9	9	9	9
D3_017_max	5	4	4	9	9	9
D3_017_min	7	4	4	9	5	5
D3_107_max	3	3	3	9	9	9
D3_107_min	6	3	3	8	5	8
H1_083_max	9	9	9	9	5	9
H1_083_min	4	3	3	9	4	4
H1_123_max	7	3	3	9	7	7
H1_123_min	5	3	3	9	7	7
Rho_175_max	5	3	3	9	8	8
Rho_175_min	7	3	3	9	6	5
Rho_149_max	9	8	8	9	8	8
Rho_149_min	7	4	4	9	8	8
ref_RB_best	4	4	4	8	8	8
ref_RB_worst	1	1	1	6	6	6
ref_LB_2000	3				-	
ref_LB_2003	3				-	

SIFt profiles were next compared with pharmacophore features map to retrieve complementary interaction points.

In addition, a fast automatic algorithm was prepared to search a small collection of pharmacophore models, which linear combinations showed much better performance, compared with single hypotheses. Moreover, this algorithm allowed the generation of parameter-directed combinations adjusted to maximize high recall, precision, specificity, etc. The simulations showed that optimization of MCC or the F-score produced combinations with balanced performance, while the optimization of recall was able to capture substantially more new active compounds than the best ligand-based hypotheses. The proposed method is superior to the classic ligand-based approach because it is able to combine information from the ligand and binding site topology. Including information from the receptor binding site, caused functional and spatial verifications of ligand molecule, which takes into account the chemical groups involved in the molecular recognition.

This study showed that the influence of the template selection on the final efficiency of the optimized pharmacophore ensembles was not significant and that the performance changes were similar to those created by different conformations of models built on a single template.

■ ASSOCIATED CONTENT

Supporting Information

The algorithm of matching of dipole−dipole interaction (Appendix 1), description of the obtained pharmacophore maps (Appendix 2), the developed algorithm to optimize the linear combination of pharmacophore models (Appendix 3), the used reference pharmacophore models (Appendix 4), the full results of pharmacophore query searching (Appendix 5), and the structure of centroids used in the training set (Appendix 6). This material is available free of charge via the Internet at http://pubs.acs.org.

dx.doi.org/10.1021/ci4005207 | J. Chem. Inf. Model. 2013, 53, 3233−3243

AUTHOR INFORMATION

Corresponding Author

*Phone: +48 12 662 33 65. E-mail: bojarski@if-pan.krakow.pl.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The study was partly supported by the projects: UDA-POIG.01.03.01-12-100/08-00 (http://www.prokog.pl) and UDA-POIG.01.01.02-12-004/09-00 (http://www.de-me-ter.pl), cofinanced by the European Union from the European Fund of Regional Development (EFRD).

ABBREVIATIONS

SIFt, structural interaction fingerprints; AR, aromatic group; PI, positive ionizable; HYD, hydrophobic region; HBA, hydrogen-bond acceptor; HBD, hydrogen-bond donor; DUD, directory of useful decoys

REFERENCES

(1) Schuster, D.; Maurer, E. M.; Laggner, C.; Nashev, L. G.; Wilckens, T.; Langer, T.; Odermatt, A. The discovery of new 11beta-hydroxysteroid dehydrogenase type 1 inhibitors by common feature pharmacophore modeling and virtual screening. *J. Med. Chem.* **2006**, *49*, 3454−3466.

(2) Macdougall, I. J. A.; Griffith, R. Pharmacophore design and database searching for selective monoamine neurotransmitter transporter ligands. *J. Mol. Graphics Modell.* **2008**, *26*, 1113−1124.

(3) Langer, T.; Hoffmann, R. D. *Pharmacophores and pharmacophore searches*; Mannhold, R., Kubinyi, H., Folkers, G., Eds.; WILEY-VCH Verlag GmbH & Co. KGaA: Weinheim, 2006; p 3.

(4) Sanders, M. P. A.; Verhoeven, S.; de Graaf, C.; Roumen, L.; Vroling, B.; Nabuurs, S. B.; de Vlieg, J.; Klomp, J. P. G. Snooker: A structure-based pharmacophore generation tool applied to class A GPCRs. *J. Chem. Inf. Model.* **2011**, *51*, 2277−2292.

(5) Klabunde, T.; Giegerich, C.; Evers, A. Sequence-derived three-dimensional pharmacophore models for G-protein-coupled receptors and their application in virtual screening. *J. Med. Chem.* **2009**, *52*, 2923−2932.

(6) Wang, R.; Gao, Y.; Lai, L. LigBuilder: A multi-purpose program for structure-based drug design. *J. Mol. Model.* **2000**, *6*, 498−516.

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

(8) Wolber, G.; Langer, T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* **2005**, *45*, 160−169.

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

(10) Goodford, P. J. Computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849−857.

(11) 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.

(12) Meslamani, J.; Li, J.; Sutter, J.; Stevens, A.; Bertrand, H.-O.; Rognan, D. Protein-ligand-based pharmacophores: generation and utility assessment in computational ligand profiling. *J. Chem. Inf. Model.* **2012**, *52*, 943−955.

(13) *Discovery Studio*, version 3.1.0; Accelrys: San Diego, CA, 2012.

(14) Rogers, D.; Hopfinger, A. J. Application of genetic function approximation to quantitative structure−activity relationships and quantitative structure−property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.

(15) Salam, N. K.; Nuti, R.; Sherman, W. Novel method for generating structure-based pharmacophores using energetic analysis. *J. Chem. Inf. Model.* **2009**, *49*, 2356−2368.

(16) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, 1100−1107.

(17) *Canvas*, version 1.4; Schrödinger, LLC, New York, NY, 2011.

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

(19) Irwin, J. J.; Shoichet, B. K. ZINC − A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.

(20) Beutler, T. C.; Dill, K. A. A fast conformational search strategy for finding low energy structures of model proteins. *Protein Sci.* **1996**, *5*, 2037−2043.

(21) The UniProt Consortium, Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2012**, *40*, D71−D75.

(22) Amico, M.; Finelli, M.; Rossi, I.; Zauli, A.; Elofsson, A.; Viklund, H.; von Heijne, G.; Jones, D.; Krogh, A.; Fariselli, P.; Luigi Martelli, P.; Casadio, R. PONGO: A web server for multiple predictions of all-alpha transmembrane proteins. *Nucleic Acids Res.* **2006**, *34*, 169−172.

(23) Sigrist, C. J. A.; Cerutti, L.; Hulo, N.; Gattiker, A.; Falquet, L.; Pagni, M.; Bairoch, A.; Bucher, P. PROSITE: A documented database using patterns and profiles as motif descriptors. *Briefings Bioinf.* **2002**, *3*, 265−274.

(24) Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **1999**, *292*, 195−202.

(25) Šali, A.; Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **1993**, *234*, 779−815.

(26) Kołaczkowski, M.; Nowak, M.; Pawłowski, M.; Bojarski, A. J. Receptor-based pharmacophores for serotonin 5-HT$_7$R antagonists-implications to selectivity. *J. Med. Chem.* **2006**, *49*, 6732−6741.

(27) *Glide*, version 5.7; Schrödinger, LLC: New York, NY, 2011.

(28) *Epik*, version 2.7; Schrödinger, LLC: New York, NY, 2011.

(29) Deng, Z.; Chuaqui, C.; Singh, J. Structural interaction fingerprint (SIFt): A novel method for analyzing three-dimensional protein-ligand binding interactions. *J. Med. Chem.* **2004**, *47*, 337−344.

(30) Mordalski, S.; Kosciolek, T.; Kristiansen, K.; Sylte, I.; Bojarski, A. J. Protein binding site analysis by means of structural interaction fingerprint patterns. *Bioorg. Med. Chem. Lett.* **2011**, *21*, 6816−6819.

(31) Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442−451.

(32) Davis, J.; Goadrich, M. The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd international conference on Machine learning - ICML '06* **2006**, 233−240.

(33) McGregor, M. J.; Muskal, S. M. Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 569−574.

(34) McGregor, M.; Muskal, S. M. Pharmacophore fingerprinting. 2. Application to primary library design. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 117−125.

(35) Sud, M. MayaChemTools: An open source package for computational discovery. In *243rd ACS National Meeting & Exposition, March 25−29 2012, San Diego, CA*, 2012.

(36) *JChem*, version 5.4.1; ChemAxon Kft.: Budapest, Hungary, 2011.

(37) Lopez-Rodriguez, M. L.; Porras, E.; Benhamu, B.; Ramos, J. A.; Morcillo, J. M.; Lavandera, J. L. First pharmacophoric hypothesis for 5-HT$_7$ antagonism. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1097−1100.

(38) Lopez-Rodriguez, M. L.; Porras, E.; Morcillo, J. M.; Benhamu, B.; Soto, L. J.; Lavandera, J. L.; Ramos, J. A.; Olivella, M.; Campillo, M.; Pardo, L. Optimization of the pharmacophore model for 5-HT$_7$R

antagonism. Design and synthesis of new naphtholactam and naphthosultam derivatives. *J. Med. Chem.* **2003**, *46*, 5638−5650.
(39) Bojarski, A. J. Pharmacophore models for metabotropic 5-HT receptor ligands. *Curr. Top. Med. Chem.* **2006**, *6*, 2005−2026.