Property-Based Design of GPCR-Targeted Library

Konstantin V. Balakin,* Sergey E. Tkachenko, Stanley A. Lang, Ilya Okun, Andrey A. Ivashchenko, and Nikolay P. Savchuk

Chemical Diversity Labs, Inc., 11575 Sorrento Valley Road, San Diego, California 92121

Received May 15, 2002

The design of a GPCR-targeted library, based on a scoring scheme for the classification of molecules into "GPCR-ligand-like" and "non-GPCR-ligand-like", is outlined. The methodology is a valuable tool that can aid in the selection and prioritization of potential GPCR ligands for bioscreening from large collections of compounds. It is based on the distillation of knowledge from large databases of GPCR and non-GPCR active agents. The method employed a set of descriptors for encoding the molecular structures and by training of a neural network for classifying the molecules. The molecular requirements were profiled and validated by using available databases of GPCR- and non-GPCR-active agents [5736 diverse GPCR-active molecules and 7506 diverse non-GPCR-active molecules from the Ensemble Database (Prous Science, 2002)]. The method enables efficient qualification or disqualification of a molecule as a potential GPCR ligand and represents a useful tool for constraining the size of GPCR-targeted libraries that will help speed up the development of new GPCR-active drugs.

INTRODUCTION

It is now well understood that the path between a discovery of a compound with high affinity binding to a target and subsequent developing of a successful drug can be very protracted or even may not be realized at all. Poor absorption and related poor pharmacokinetics are one of the main reasons for attrition in the drug development process. It is now widely recognized that physicochemical, pharmacokinetic, and biopharmaceutical properties need to be addressed early in the drug discovery process. Several papers have discussed the thesis that drugs have distinct properties differentiating them from other chemicals. Using artificial neural network analysis, a compound can be predicted as being "drug-like" or "non-drug-like".[1,2] Similarly, in a study on CNS-active drugs that used a neural networks analysis, CNS-active drugs could be distinguished from ones without CNS activity.[3] Probably, the best-known method of computational prediction of intestinal absorption of a compound is the "rule of five"[4] devised by Lipinski and co-workers at Pfizer (Groton, CT) from an analysis of 2245 drugs from the WDI believed to have entered Phase II trials. The "rule of five" generates an alert (indicating possible absorption problems) for compounds, where any two of the following conditions are met:

- Molecular weight > 500.
- Number of hydrogen-bond acceptors > 10.
- Number of hydrogen-bond donors > 5.
- Calculated logP > 5.0 (if ClogP is used) or > 4.15 (if MlogP is used).

Computation of these related to ADME (Absorption, Distribution, Metabolism, Excretion) properties is now available in a number of commercial software packages. The "rule-of-five" should be seen as a qualitative absorption/permeability predictor, rather than a quantitative predictor.[5] A distribution of molecular properties in drug related chemical databases has been studied as another approach to understand the "drug-like" or "lead-like" characteristics of small molecules.[6−9] All these aforementioned analyses point to a critical combination of physicochemical and structural properties, which to a large extent can be manipulated by a medicinal chemist, that are required for each analysis. Recently, the use of these tools in medicinal chemistry had been termed by van de Waterbeemd et al.[10] as a "property-based design". Regarding these properties, the authors intended that physicochemical as well as pharmacokinetic characteristics of compounds were taken into consideration. These properties have been neglected for a long time by medicinal chemists, who in many cases considered the strongest receptor binding values as the ultimate goal.

In this paper, we present a methodology of the property-based design of a GPCR-targeted compound library. The superfamily of seven-transmembrane-domain G-protein-coupled receptors (GPCRs) is the diverse group of transmembrane proteins involved in signal transduction.[11,12] GPCRs initiate a cascade of cellular responses to diverse extracellular mediators and nearly 50% of marketed drugs act through modulation of the GPCR functions. In addition, for the GPCRs, for which a ligand has not yet been identified (orphan GPCRs), this methodology could provide a path to discover new cellular components that are important in human physiology.[13] The process of characterizing and thus removing these novel proteins from the orphan GPCR list could assist in accelerating research in human physiology and pharmacology.

Our goal is to construct an algorithm utilizing simple, automated procedures for designing combinatorial libraries that would show preferential GPCR-activity. "GPCR-activity" is assumed here as the ability of a compound to be a successful ligand for a GPCR. If a quantitative relationship,

* Corresponding author phone: (858)794-4860; fax: (858)794-4931; e-mail: kvb@chemdiv.com.

10.1021/ci025538y CCC: $22.00
Published on Web 10/31/2002

Table 1. Reference Database of GPCR-Ligands

no.	GPCR targets	ligand type	no. of compds
1	5-HT1A receptor	agonists	899
2	5-HT1D receptor	agonists and antagonists	314
3	5-HT2B receptor	antagonists	64
4	α1/α2-adrenoceptor	antagonists	213
5	β-adrenoceptor	agonists and antagonists	684
6	bradykinin B2 receptor	agonists and antagonists	44
7	cannabinoid CB1 and CB2 receptors	antagonists	10
8	δ-opioid receptor	agonists	172
9	dopamine autoreceptor	modulators	165
10	dopamine D1 receptor	antagonists	190
11	dopamine D2 receptor	agonists	855
12	dopamine D3 receptor	antagonists	231
13	dopamine D4 receptor	antagonists	14
14	endothelin ETA/ETB receptor	antagonists	98
15	κ-opioid receptor	agonists	155
16	μ-opioid receptor	agonists and antagonists	64
17	muscarinic receptor	agonists	94
18	neuropeptide Y receptor	antagonists	21
19	oxytocin receptor	antagonists	65
20	sigma-receptor	antagonists	363
21	tachykinin NK1 receptor	antagonists	783
22	tachykinin NK2 receptor	antagonists	128
23	vasopressin V1/V2 receptor	antagonists	166
	total no. of compds in training set[a]		5736

[a] The total number of compounds is not equal to the sum of the shown values, as some compounds are not selective and manifest activity against more than one target.

based on the thorough assessment of a contribution values of several important physicochemical parameters influencing GPCR activity, could be successfully established, it would not only permit the estimation of potential "GPCR-ligand-likeness" of candidate compounds prior to synthesis but would also provide information concerning the modification of the structural features necessary for GPCR-activity. In this work, we have attempted to develop an effective scoring system for the classification of molecules into GPCR-actives and GPCR nonactives using the neural network classification approach.

METHODS

Databases. 5736 known GPCR ligands belonging to 23 different GPCR classes (Table 1) were used as a positive training set. For comparison, a subset of 7506 compounds, representing over 100 various non-GPCR activities (Table 2) were used as a negative training set. Compounds in the stages of (pre)clinical trials, marketed drugs, and compounds with proven in vivo activity were included into these two data sets. All compounds were selected from the Ensemble database[14] which is a licensed database of known pharmaceutical agents compiled from the patent and scientific literature. Structures were extracted according to the assigned activity class, where the class indicates a single protein target or a therapeutic area. Some structures have one or more key words in the "mechanism of action" and "therapeutic group" fields. We assumed that a molecule is GPCR-active, only if it contains the indication on a GPCR protein in the "mechanism of action" field. All other compounds were considered as GPCR-inactive. For therapeutic groups for which a biotarget was not clearly indicated or multiple biotargets were reported, we included into the GPCR(−) data set only compounds with a mechanism of action unrelated to GPCRs (for example, oncolytic drugs with antimitotic action). There are some unavoidable limitations to using databases such as Ensemble. Since not every non-GPCR-active compound has been tested for GPCR-activity, one cannot assume that a compound without a particular key word is GPCR-inactive. Thus there are probably a number of false inactives. Molecules were filtered based on molecular weight range (200−650) and atom type content (only C, N, O, H, S, P, F, Cl, Br, and I allowed).

Diversity parameters for the positive, GPCR(+), and negative, GPCR(−), reference compound sets are shown in Table 3. As evident from the number of screens, the number of unique heterocyclic fragments, and the diversity coefficients (all these parameters are calculated using ChemoSoft software tool[15,16]), each of the two compound databases has a high level of diversity and can be considered as a good representation of the GPCR-active and non-GPCR-active compounds. The diversity parameters for the GPCR(−) database are naturally higher, as the compounds it contains are active ligands to a considerably larger number of diverse biotargets.

Descriptors. A set of eight descriptors that was calculated from 2D representations of molecules was explored. The evaluated descriptors are listed in Table 4. All descriptors were calculated using the CDL proprietary software tool, ChemoSoft.[15,16] For calculations of $LogD_{74}$, $LogS_w$, and FA, the SLIPPER program,[17,18] integrated into the ChemoSoft environment, was used. These ADME-related properties of molecules are highly significant in the context of pharmacokinetic characteristics of drug candidates and are important parameters contributing to the predictive power of the developed model.

Statistical Analysis. t′-Statistics has been used to measure the ability of molecular descriptors to discriminate between the two categories of compounds, GPCR(+) and GPCR(−). t′-Test is a kind of a two-tailed heteroscedastic t-test that is convenient for comparative evaluation of large normally distributed samples. It is frequently used to determine whether two sample means are equal or different if the population variances are unequal. The difference in the means of the distributions is expressed in accordance with the formula

$$t' = (X1 - X2)/\sqrt{(\sigma_1^2/n_1 + \sigma_2^2/n_2)}$$

where for each of the two compound sets, GPCR(+) and GPCR(−), X is the mean, σ^2 is the variance, n is the sample size, and 1 and 2 denote the corresponding set. For the large samples studied in this work, there is a trend toward a normal distribution. The only exception is distribution of the number of H-bond donors which is non-normal in the population (Figure 1). However, as the sample size (of samples used to create the sampling distribution of the mean) increases, the shape of the sampling distribution becomes normal. For $n = 20-30$, the shape of that distribution is almost perfectly normal. Therefore, according to the central limit theorem, t′-statistics is applicable to test the significance of the difference between the two distributions for each descriptor (Table 4). The t′-value was not calculated for FA, as the

Table 2. Reference Base of Non-GPCR Ligands[14] Used in This Study

no.	ligand type	no.	%	no.	ligand type	no.	%
	Compounds with Assignment to Particular Biotarget: No. 4269; 56.9%						
1	3-hydroxyanthranilate 34-dioxygenase inhibitors	7	0.09	49	lanosterol 14alpha-demethylase inhibitors	5	0.07
2	ACAT inhibitors	34	0.45	50	lanosterol synthase inhibitors	41	0.55
3	ACE inhibitors	85	1.13	51	LDL-receptor up-regulators	14	0.19
4	AGE inhibitors	38	0.51	52	lipoxygenase inhibitors	9	0.12
5	AIT Tat inhibitors	6	0.08	53	MAO inhibitors	43	0.57
6	alanine racemase inhibitors	1	0.01	54	matrix metalloproteinase inhibitors	169	2.25
7	aldose reductase inhibitors	270	3.60	55	microsomal triglyceride transfer protein inhibitors	30	0.40
8	alpha-glucosidase inhibitors	27	0.36	56	microtubule inhibitors	70	0.93
9	alpha-mannosidase inhibitors	4	0.05	57	Na/H exchange inhibitors	38	0.51
10	AMPA ligands	9	0.12	58	NAALADase inhibitors	23	0.31
11	antiestrogens	46	0.61	59	NAD/ADF-ribosyltransferase inhibitors	21	0.28
12	antithyroids	1	0.01	60	neutral endopeptidase inhibitors	29	0.39
13	ATPase inhibitors	23	0.31	61	NF-kappaB inhibitors	9	0.12
14	beta-amyloid protein neurotoxicity inhibitors	14	0.19	62	nicotinic agonists	18	0.24
15	calcium channel blockers/openers	60	0.80	63	nitric oxide donors	2	0.03
16	calmodulin antagonists	4	0.05	64	nitric oxide synthase inhibitors	8	0.11
17	calpain inhibitors	22	0.29	65	NMDA ligands	66	0.88
18	cathepsin inhibitors	24	0.32	66	norepinephrine reuptake inhibitors	4	0.05
19	cholesterol esterase inhibitors	30	0.40	67	nuclear receptor ligands (retinoids)	121	1.61
20	citrate lyase inhibitors	2	0.03	68	nuclear receptor ligands (vitamin D analogues)	113	1.51
21	CMP-KDO synthase inhibitors	3	0.04	69	PAF antagonists	17	0.23
22	collagenase inhibitors	71	0.95	70	phosphodiesterase inhibitors	66	0.88
23	COMT inhibitors	8	0.11	71	phospholipase inhibitors	76	1.01
24	corticosteroids	5	0.07	72	plasmepsin inhibitors	4	0.05
25	cyclin-dependent kinase inhibitors	21	0.28	73	potassium channel modulators	78	1.04
26	cyclooxygenase inhibitors	16	0.21	74	progesterone antagonists	11	0.15
27	cysteine protease inhibitors	10	0.13	75	prolyl endopeptidase inhibitors	50	0.67
28	cytokine modulators	42	0.56	76	prolyl-4-hydroxylase inhibitors	40	0.53
29	dihydroorotate dehydrogenase inhibitors	2	0.03	77	protein kinase C inhibitors	13	0.17
30	dipeptidyl peptidase inhibitors	9	0.12	78	protein tyrosine phosphatase inhibitors	3	0.04
31	DNA polymerase inhibitors	1	0.01	79	purine-nucleoside phosphorylase inhibitors	25	0.33
32	DNA topoisomerase inhibitors	8	0.11	80	renin inhibitors	10	0.13
33	efflux pumps inhibitors	4	0.05	81	reverse transcriptase inhibitors	46	0.61
34	enkephalinase inhibitors	3	0.04	82	ribonucleoside diphosphate reductase inhibitors	3	0.04
35	estrogen agonists/antagonists	61	0.81	83	RXR ligands	112	1.49
36	factor XIIIa, VIIa, Xa inhibitors	262	3.49	84	S-adenosyl-L-homocysteine hydrolase inhibitors	22	0.29
37	farnesyl transferase inhibitors	172	2.29	85	sodium channel blockers	54	0.72
38	fibrinogen gpIIb, IIIa receptor inhibitors	450	6.00	86	squalene synthase/epoxidase inhibitors	202	2.69
39	geranyl protein transferase inhibitors	13	0.17	87	steroid 5alpha-reductase inhibitors	15	0.20
40	glucose-6-phospatase inhibitors	9	0.12	88	telomerase inhibitors	2	0.03
41	glutamate release inhibitors	19	0.25	89	thromboxane synthase inhibitors	160	2.13
42	glycogen phosphorylase inhibitors	6	0.08	90	thymidine kinase inhibitors	3	0.04
43	HCV NS3 protease inhibitors	13	0.17	91	TNF-alpha antagonists	54	0.72
44	HMG-CoA reductase inhibitors	286	3.81	92	tyrosine kinase inhibitors	65	0.87
45	inhibitors of plasminogen activators	3	0.04	93	U-PA inhibitors	5	0.07
46	interleukin antagonists	30	0.40	94	VLA-4 antagonists	15	0.20
47	kainate antagonists	21	0.28	95	xanthine oxidase inhibitors	26	0.35
48	kynureninase inhibitors	4	0.05				
	Compounds without Assignment to Particular Biotarget:[a] No. 3237; 43.1%						
96	antiarthritic drugs	25	0.33	103	dermatologic drugs	54	0.72
97	antibacterial drugs	506	6.74	104	drugs for treatment of protozoal diseases	81	1.08
98	antibiotics	290	3.86	105	gastrointestinal drugs	15	0.20
99	antidiabetic drugs	154	2.05	106	immunomodulating drugs	420	5.60
100	anti-HIV drugs	214	2.85	107	oncolytic drugs	720	9.59
101	antipsoriatic drugs	63	0.84	108	renal/urologic drugs	289	3.85
102	antiviral drugs	617	8.22		total[a]	7506	

[a] The total number of compounds is not equal to the sum of the shown values, as some compounds can be assigned to more than one therapeutic group.

Table 3. Diversity Parameters for the Reference Databases

no.	parameter	GPCR(+) database	GPCR(−) database
1	total number of compounds	5736	7506
2	number of screens[a]	8247	15770
3	number of unique heterocycles	388	1315
4	substructural diversity[b]	0.802	0.851
5	heterocycles diversity,[b] max	0.929	0.964

[a] Screens are simple structural fragments, centroids, with the topological distance equal to 1 bond length between the central atom and the atoms maximally remote from it. [b] Cosine coefficients are calculated, and the sums of nondiagonal similarity matrix elements are used in ChemoSoft program as a diversity measure.

distributions of this parameter are not normal for the studied databases.

Table 4. Descriptors Used for Analysis

descriptor	Δ[a]	t′
molecular weight[b]	7.59	4.38
log of distribution coeff in 1-octanol/ water at pH 7.4[c]	5.97	148.30
number of H-bond donors[b]	−0.73	27.80
number of H-bond acceptors[b]	−0.99	26.18
number of rotatable bonds[b]	−1.62	19.29
number of aromatic bonds[b]	2.47	24.03
log of solubility in water at pH 7.4 (expressed in g/mL)[c]	−5.16	117.56
fractional absorption[c]		

[a] Δ − difference between GPCR(+) and GPCR(−) mean values. [b] Descriptors are calculated with the ChemoSoft[15,16] software. [c] Descriptors are calculated with SLIPPER program[17,18] integrated into the ChemoSoft environment.

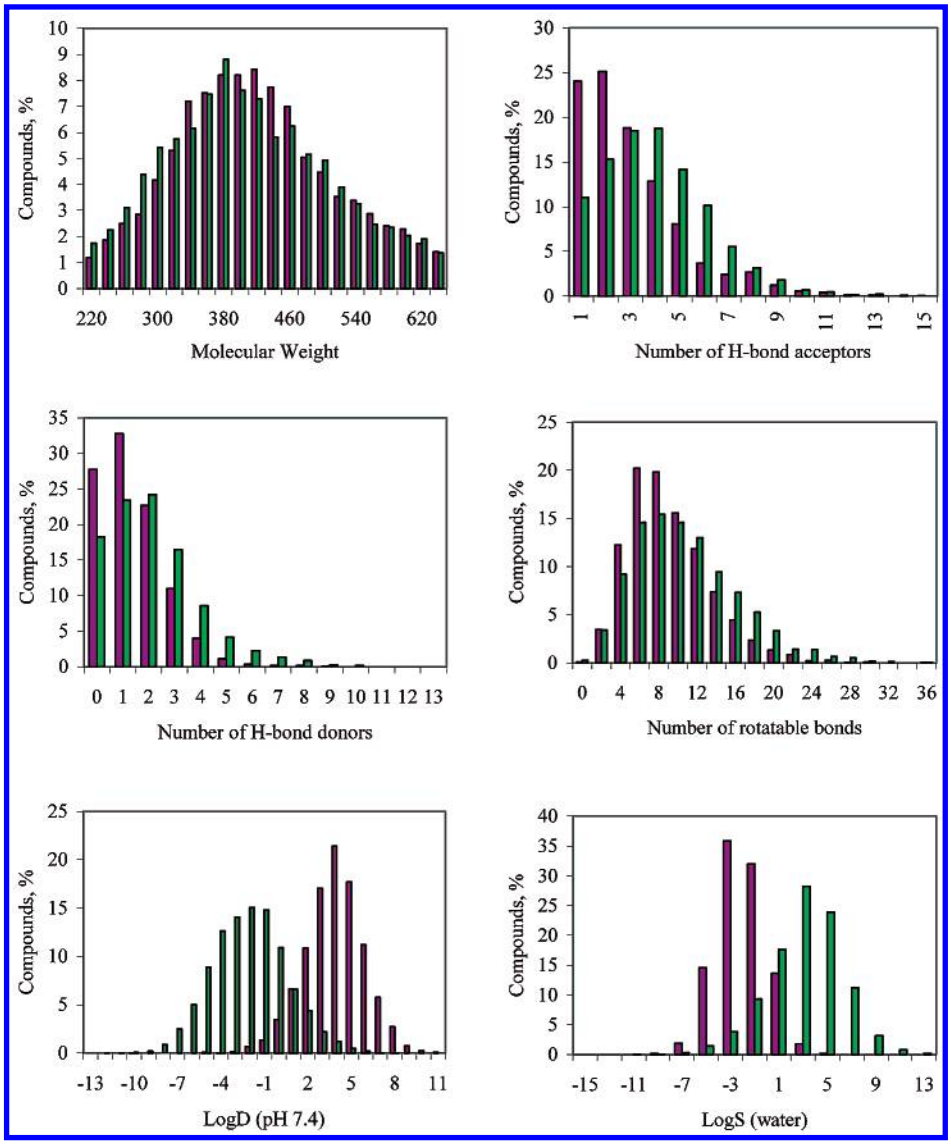

Figure 1. Property distribution profiles for data sets consisting of 5736 GPCR-active compounds (violet) and 7506 GPCR nonactive compounds (green).

Table 5. Correlation Coefficients between the Eight Descriptors of the GPCR-Actives Set, GPCR(+)

	MW	a_acc	a_don	b_ar	b_rotN	$LogD_{74}$	$LogS_w$	FA
MW	1	0.58	0.17	0.54	0.64	*0.36*	−0.37	−0.38
a_acc		1	0.17	0.18	0.47	−0.11	0.01	−0.32
a_don			1	0.30	0.23	−0.18	0.17	−0.42
b_ar				1	0.18	*0.36*	−0.32	−0.19
b_rotN					1	0.033	−0.09	−0.32
$LogD_{74}$						1	−0.70	0.19
$LogS_w$							1	***−0.19***
FA								1

Table 6. Correlation Coefficients between the Eight Descriptors of the GPCR Nonactives Set, GPCR(−)

	MW	a_acc	a_don	b_ar	b_rotN	$LogD_{74}$	$LogS_w$	FA
MW	1	0.50	0.20	0.31	0.64	*−0.33*	0.20	−0.36
a_acc		1	0.38	−0.08	0.48	0.08	−0.29	−0.43
a_don			1	0.15	0.23	0.16	−0.30	−0.47
b_ar				1	−0.14	**−0.24**	0.30	0.01
b_rotN					1	−0.09	−0.01	−0.35
$LogD_{74}$						1	−0.74	−0.32
$LogS_w$							1	***0.45***
FA								1

Training and Test Sets. For the neural network modeling, five randomizations corresponding to the complementary training/cross-validation/testing sets were considered. For the generation of neural network model, the total of 13 242 molecules of the GPCR(+) set (5736 compounds) and GPCR(−) set (7506 compounds) were randomized and subdivided into three categories: (1) training set of 6621 compounds (50% of the total number of compounds), (2) cross-validation set of 3310 compounds (25%), and (3) test set of 3311 compounds (25%). The cross-validation set was used to avoid over-training during the development of neural network models.

Neural Network. The NeuroSolution 4.0 program[19] was used for all neural network operations. Feed-forward nets were constructed that consist of six input neurons (descriptors 1−6, Table 4), one hidden layer with four processing elements, and two output neurons. The networks were trained with the molecular descriptors as input values and the scores as output values. The final score was calculated by subtraction of "GPCR-activity" score from "GPCR non-activity" score. The back-propagated nets were trained following the momentum learning rule as implemented in the NeuroSolution program. The training was performed over 1000 iterations.

Table 7. Results of the Test Set Classification with the Neural Net Model for Five Independent Randomizations

	test compound sets											
	GPCR(+) randomizations						GPCR(−) randomizations					
prediction	1	2	3	4	5	mean	1	2	3	4	5	mean
number of GPCR(+) predicted	1326	1313	1313	1319	1326	1319	155	136	138	139	128	139
number of GPCR(−) predicted	119	119	139	122	129	126	1711	1743	1721	1731	1728	1727
percent of correct predictions	91.8	91.7	90,0	91.5	91.1	91.2	91.7	92.8	92.6	92.6	93.1	92.6

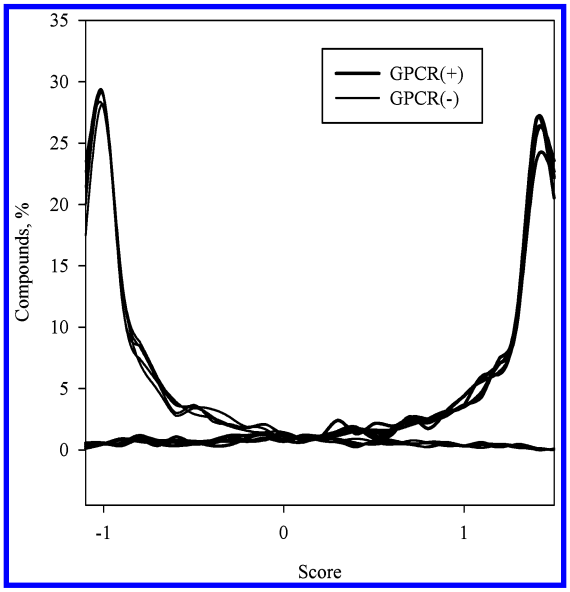

Figure 2. Compound distributions on the scale of prediction scores for the test set. The data are shown for five independent randomizations.

RESULTS

Molecular Descriptors. A number of studies have suggested that limited sets of molecular descriptors are sufficient to capture important molecular features. Limited descriptor sets have successfully been used to cluster compound collections, study their diversity, and predict their biological properties.[20−22] It was shown that a set of special structural fragments, or keys, and several additional 1D/2D descriptors accounting for aromatic character, hydrogen-bonding capacity, and molecular flexibility are usually sufficient to effectively partition compounds according to biological activity. In this work, we show that the use of a simple set of physicochemical descriptors for creating the neural network classification method enables effective discrimination between GPCR-active and GPCR nonactive compounds.

The set of descriptors listed in Table 4 was utilized. The descriptors 1−6 were used in neural network modeling experiments. The other two descriptors ($LogS_w$, FA) were used as additional assessment parameters for the models developed.

Differences between GPCR(+) and GPCR(−) Ligands. To uncover these differences, two types of analyses on the data setets of GPCR-active, GPCR(+), and GPCR nonactive, GPCR(−), compounds were performed.

The first analysis, illustrated in Figure 1, is based on a comparison of the physicochemical parameters. When comparing non-GPCR with GPCR-ligands (Figure 1, Table 4), a clear difference in some parameters is observed. Though statistically significant, there is a negligible difference (ca. 2%) in mean molecular weight among the two sets. A definitively lower number of the hydrogen-bond acceptors and donors ($\Delta a\_acc = -0.99$, $t' = 26.18$; $\Delta a\_don = -0.73$, $t' = 27.8$) as well as the number of rotatable bonds ($\Delta b\_rot = -1.6$, $t' = 19.29$) and significantly higher number of aromatic bonds ($\Delta b\_ar = +2.5$, $t' = 24.03$) and higher lipophilicity ($\Delta LogD_{74} = +6$, $t' = 148.3$) is characteristic of the GPCR(+) ligands.

The differences between GPCR-ligands and other therapeutic agents can be, therefore, expressed as follows: GPCR-ligands, having, on average, the similar molecular weight, are less flexible (lesser number of rotatable bonds and higher number of aromatic bonds), less polar (lesser number of H-bond donors and acceptors), and distinctly more hydrophobic (larger $LogD_{74}$, lower $LogS_w$).

The second approach is based on the correlation analysis for the eight descriptors described above. The correlation matrixes for the databases of GPCR-actives and nonactives are shown in Tables 5 and 6, respectively.

Correlation analysis allows postulation of general rules for both compound sets. For example, lipophilicity and water solubility parameters of the two compound sets have an expressed negative correlation (−0.7 for GPCR(+) and −0.74 for GPCR(−)), which should be expected from the physicochemical nature of these two parameters reflecting to opposite phenomena, lipophilicity/hydrophilicity. For the two compound sets, the molecular weight, MW, has a noticeable positive correlation with the number of hydrogen-bond acceptors (0.58 and 0.50), number of rotatable bonds (0.64 for both), and the number of aromatic bonds (0.54 and 0.31). However, the correlation of MW with the number of hydrogen-bond donors is insignificant (0.17 and 0.20) for both compound sets. Negative correlation of MW and predicted FA value is observed and is practically the same for both compound sets (−0.38 and −0.36). This negative correlation is in an agreement with the Lipinski rule stating that the penetration of larger molecules through cell membranes is usually more problematic. A moderately negative (in the range of $-0.32 \div -0.47$) correlation of such an important parameter as fractional absorption, FA, with the descriptors such as MW, the number of hydrogen-bond acceptors, a_acc, and donors, a_don, and the number of rotatable bonds, b_rotN, is similar for both compound sets. On the other hand, the lipophilicity of the compounds in each of the databases does not correlate with the number of hydrogen-bond acceptors, donors, and the number of rotatable bonds.

Tables 5 and 6 above reveal some interesting differences between the property correlations in these two, GPCR(+) and GPCR(−), compound sets. (1) Among the GPCR(+) compounds, an increase in MW has a positive correlation trend with the lipophilicity, $logD_{74}$: the larger mass of the GPCR-ligands, the higher lipophilicity of the molecule. Among the active compounds unrelated to the GPCR target,

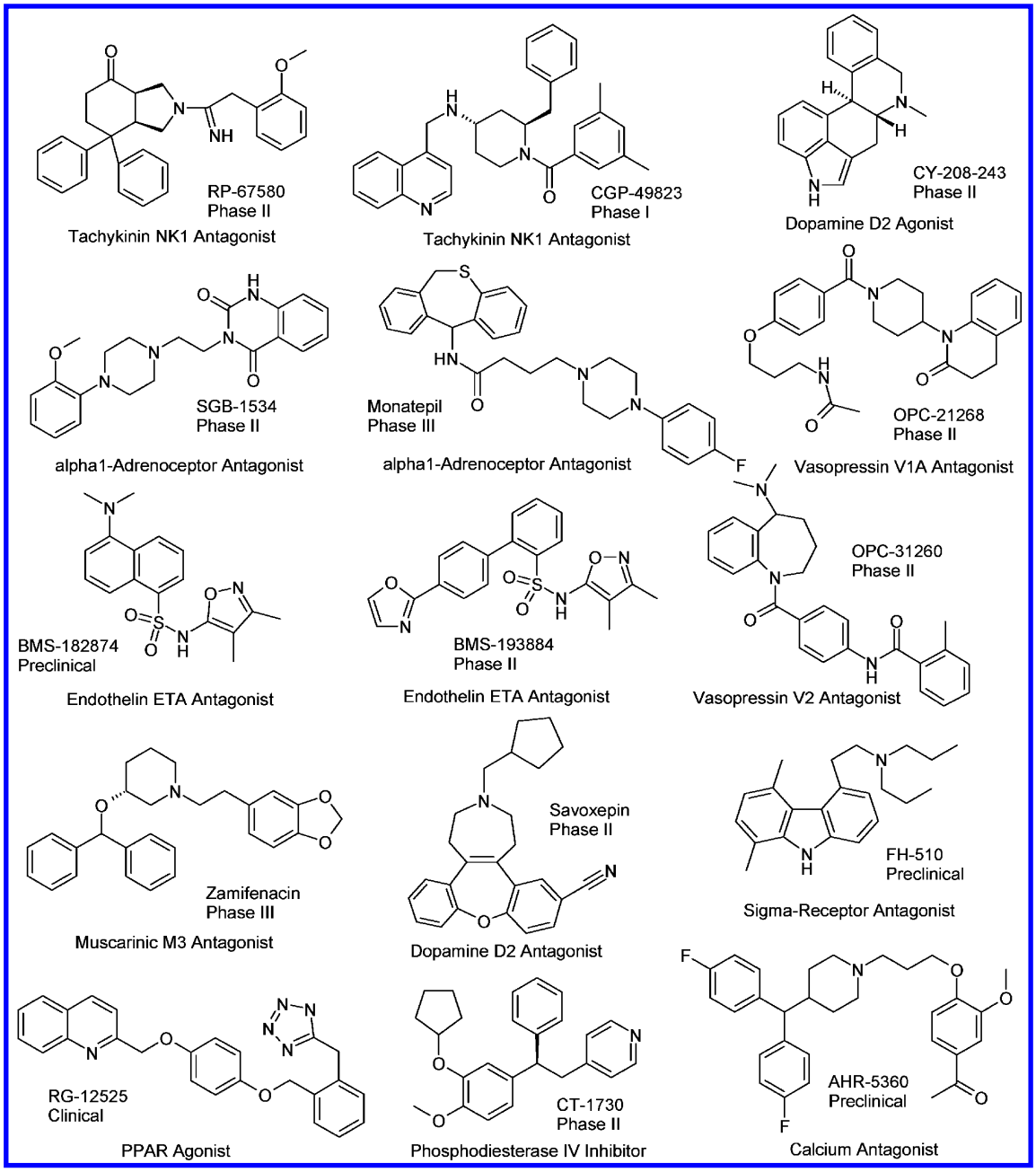

Figure 3. Examples of high-scoring compounds (score value greater than 1).

the correlation trend between the molecular weight and the lipophilicity is opposite, the higher molecular weight, the lower lipophilicity (hence higher water solubility) (italicized numbers in the tables). (2) Correlation between the number of aromatic bonds, b_ar, in the molecule and its lipophilicity, $LogD_{74}$, also has opposite trends between the two compound sets (boldface numbers in the tables). The lipophilicity of the aromatic rings of the GPCR(+) compounds becomes higher with the increase in the number of aromatic rings, whereas the opposite trend is observed for the non-GPCR compounds. This may be the result of the fact that the GPCR-(+)-ligands possess aromatic rings with higher lipophilicity than those of the non-GPCR compounds. (3) Interestingly, the predicted values of the fractional absorption, FA (italicized and boldface numbers in the tables), have a positive correlation with the water solubility, $LogS_w$, for active non-GPCR ligands (Table 6) and negative correlation with water solubility for the GPCR-active agents (Table 5).

The above observations indicate the presence of a combination of some specific physicochemical features that differentiate the GPCR-ligands from the compounds belonging to other target-specific classes.

Neural Network Development. Linear regression techniques have long been used to develop QSAR models. But these methods sometimes result in QSAR models exhibiting variability when trained with noisy data. In addition, traditional techniques often require subjective decisions to be made when the likely nonlinear relationships between structure and activity are observed. It is important that QSAR methods be quick, give unambiguous models, not rely on any subjective decisions about the functional relationships between structure and activity, and be easy to validate. Recently, methods based on neural networks have been shown to overcome some of these problems as they can account for nonlinear SARs and also can deal with linear dependencies which sometimes appear in real SAR problems.

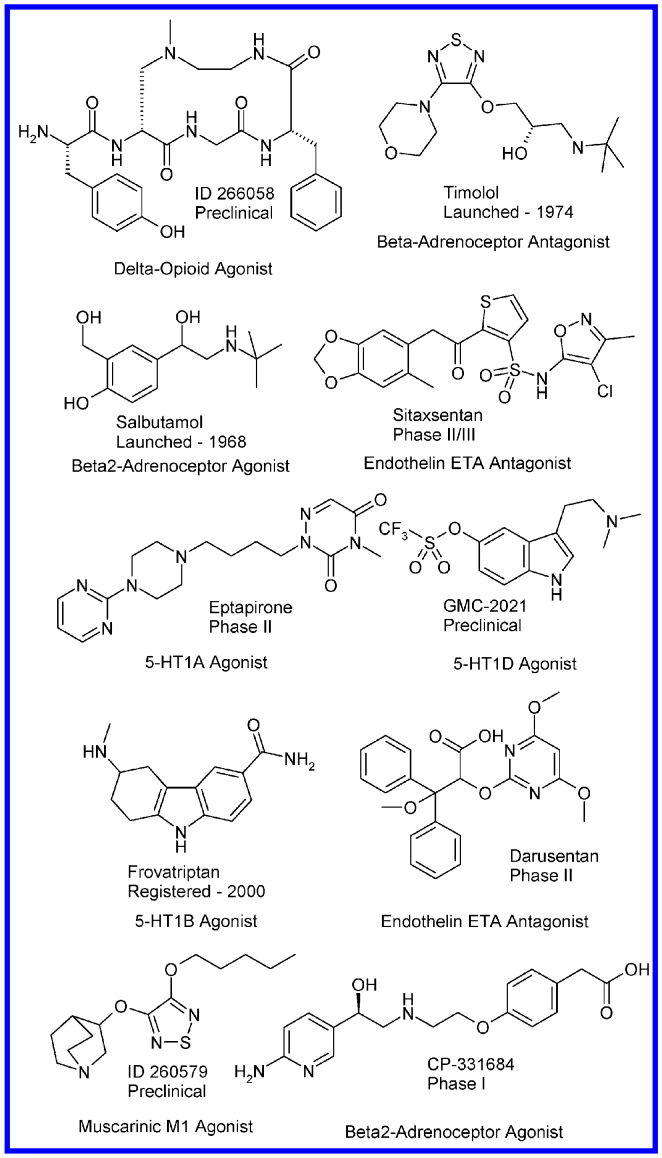

Figure 4. Examples of misclassified GPCR-active compounds (score value less than −1).

Several papers appeared that described the successful employment of different neural network approaches to distinguish different categories of compounds. Ajay et al.[2] used a Bayesian neural network for discriminating drugs from nondrugs. The network was trained using a random partition of 3500 compounds, each from the CMC and ACD databases. Two kinds of descriptors were used: a set of seven one-dimensional and 166 two-dimensional descriptors. The program was able to correctly classify 90% of the CMC compounds and misclassified only 10% of the ACD molecules. The method was demonstrated as a general method by the program's ability to correctly classify 80% of the compounds from the MDDR. The same group of researchers[3] described a solution to designing a CNS-active library based on a similar neural network classification procedure. In a database of 275 compounds, where each compound had known CNS activity, accuracy of 92% and 71%, respectively, was achieved between the actives and nonactives. Appearing back-to-back with the Ajay et al.[2] article was a contribution from Sadowski and Kubinyi,[1] in which they developed a feed-forward neural network method for discriminating drugs from nondrugs. A total of 38 416 molecules from the WDI database, as the drug set, and 169 331 molecules from the ACD, as the nondrug set, were used. The program was able to correctly classify 83% of the ACD compounds and 77% of the WDI compounds.

In the current study, the possibility of developing a neural network model that can be applied to the assessment of "GPCR-ligand-likeness" of large compound databases was examined. The learning model constructed was further used as a kind of target-specific filter to examine ca. 40 000 analogues (topological, bioisosteric, similarity) of known GPCR-ligands selected from CDL's corporate stock-available collection.

Model Development. Five independent randomization-training-testing experiments were carried out using the Generalized Feed-forward Network, a special class of Multilayer Perceptrons (MLP). The prediction quality was approximately the same for each of these five independent cycles. In control experiments with random selections of GPCR(+) and GPCR(−) test compounds, up to 92% of GPCR-ligands and 93% of non-GPCR-ligands (Table 7) were correctly predicted in the corresponding compound sets.

Figure 2 shows the distributions of the prediction scores for the test set calculated for five independent randomizations. The thick line is assigned to the GPCR-ligands, and the thin line is assigned to the non-GPCR ligands present within the test set. The separation threshold is set to the score value equal to 0 (an equal probability for a compound to be considered as either GPCR-ligand or non-GPCR-ligand). There is a clear discrimination between the GPCR-actives and GPCR nonactives in this graph. The structures with positive and negative scores were considered, respectively, as GPCR-active and GPCR nonactive predictions in Table 7.

The distinctively different distribution of the compounds with known GPCR-activity and compounds with other pharmacological activities in accordance with their scores demonstrates the high discriminative power of the trained network. As can be seen from Figure 2, the prediction accuracy of the GPCR-actives could be improved even further by setting a cutoff score to more positive values.

As an example of structures that score highly in the developed neural network model, 15 molecules with a score greater than 1 are presented. Figure 3 shows 12 GPCR-active agents and three compounds that belong to three different activity classes. All these compounds represent active leads entering into (pre)clinical trials. This figure provides some clue to the kinds of molecules the network deems to be "GPCR-ligand-like".

Misclassified Compounds. To give examples of GPCR-active compounds that were not classified correctly (false negatives), 10 arbitrary compounds with a score of less than −1 are presented (Figure 4). This gives an impression of where the limitations of the developed methodology are. It can be seen that these compounds, in general, have many highly hydrophilic and easily ionizable groups, increased number of H-bond donors and acceptors, and reduced number of aromatic atoms. As we have demonstrated above, these features are uncharacteristic of most of the GPCR-active agents. This figure gives snapshot of compounds the network assigns to the "non-GPCR-ligand-like" set.

Design of a GPCR-Active Focused Library. We have applied the developed neural network model to the generation

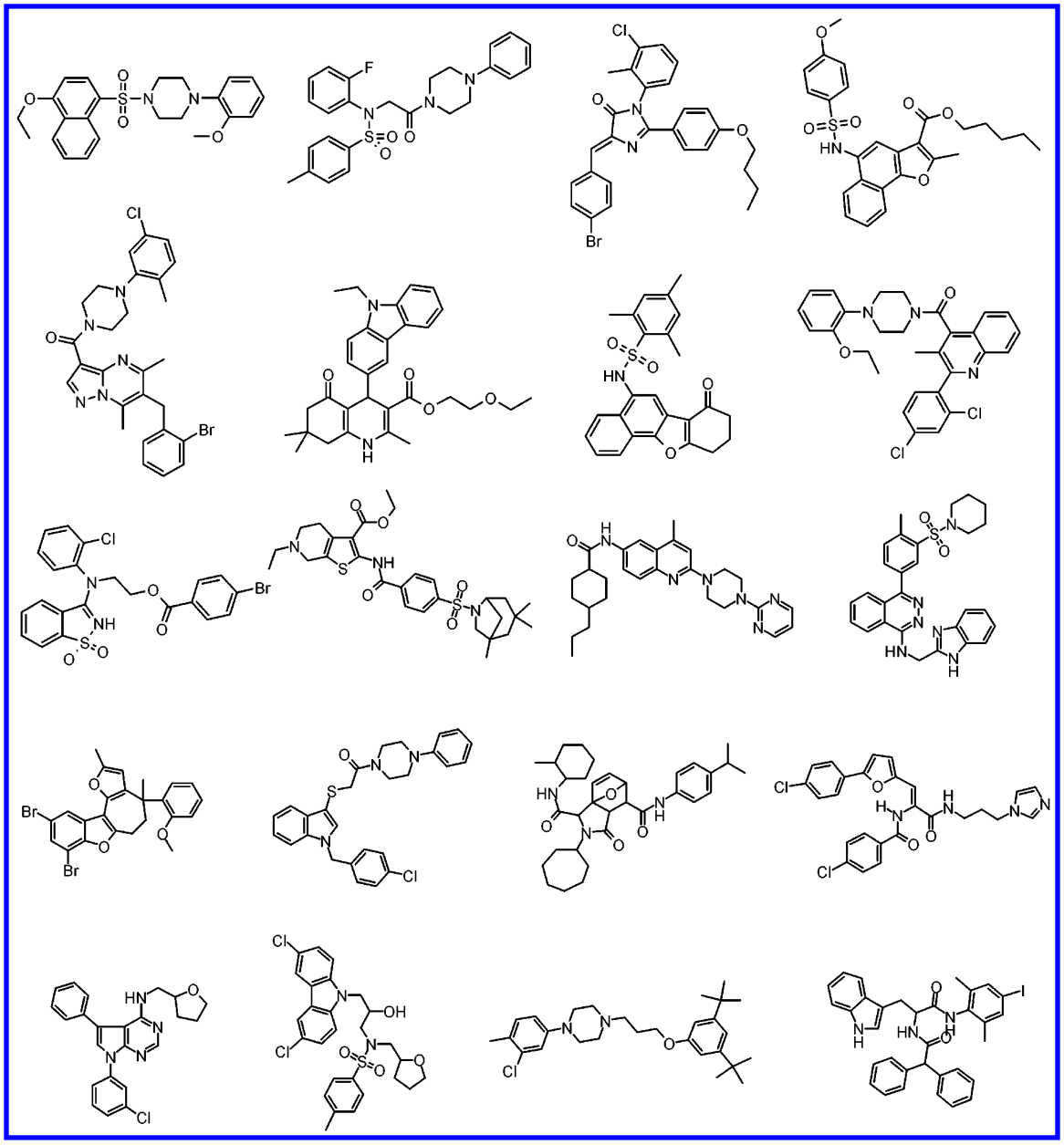

Figure 5. Examples of highest scoring structures (score value 1.44) selected from the GPCR-targeted library.

of a "GPCR-ligand-like" focused library. For this, using ChemoSoft structural similarity search engine,[16] we initially selected 40 000 compounds from the more than 180 000 CombiLab compound collection designed around 517 proprietary CDL CombiLab templates. The structural similarity was calculated relative to the set of known GPCR ligands (see Table 1). These 40 000 preselected compounds were scored and classified using the developed neural net procedure. As a result, about 30 000 compounds with positive scores were selected for inclusion into the final focused library.

In Figure 5, 20 molecules with maximal scores out of the 30 000 in the library are shown. The properties of these compounds are similar to the properties of the high-scoring molecules from the training set (see Figure 3). Figure 6 shows compounds with a score of less than −1 from the tested 40 000-compound database. Distinctive features of these compounds are a large number of highly hydrophilic and ionizable groups and lesser aromaticity as compared with the positively scored molecules. It should be noted that the described procedure is amenable to the generation of virtual GPCR-focused combinatorial libraries, and, hence, can guide *de novo* design of the focused library prior to its synthesis.

Optimization of Library Properties. We have analyzed the two sets of compounds obtained with the neural network selection, positive and negative scoring, and compared them with the set of known GPCR-ligands using simple descriptor characterization. The diagrams below (Figure 7) depict the molecular property distribution profiles obtained for the three compound sets: (i) compounds with neural net negative scores (blue line), (ii) compounds with neural net positive scores (pink line), and (iii) known GPCR-active compounds (orange line).

It is clearly evident that among the compounds selected for structural similarity to known GPCR-ligands, the descriptor distributions of the compounds with neural net positive scores are much closer to the corresponding distributions of the real life active compounds than those within the negative

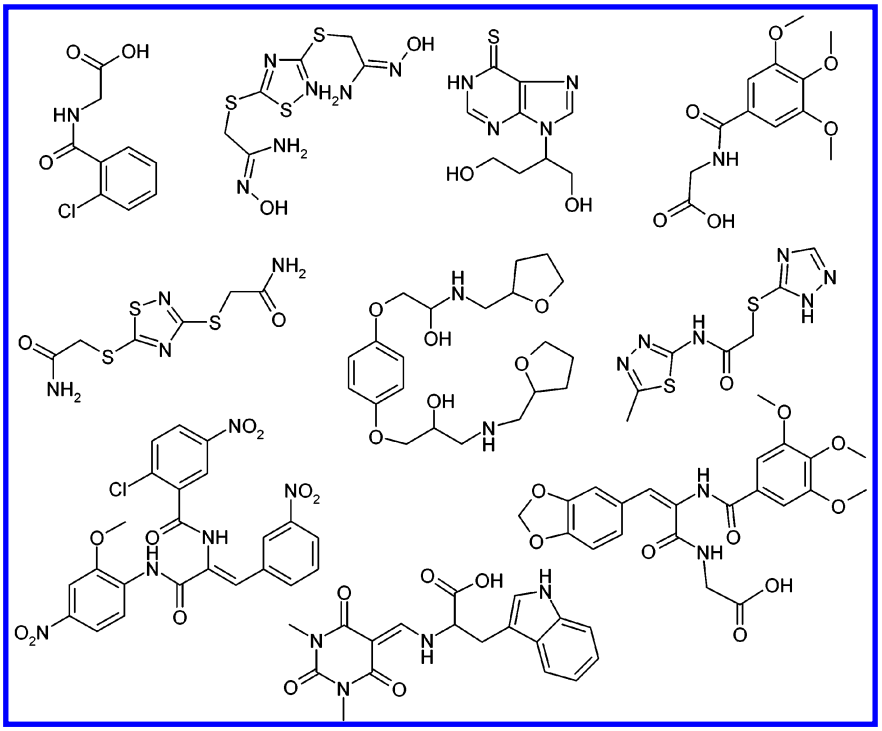

Figure 6. Examples of compounds from the 40 000-compound database (score value less than −1).

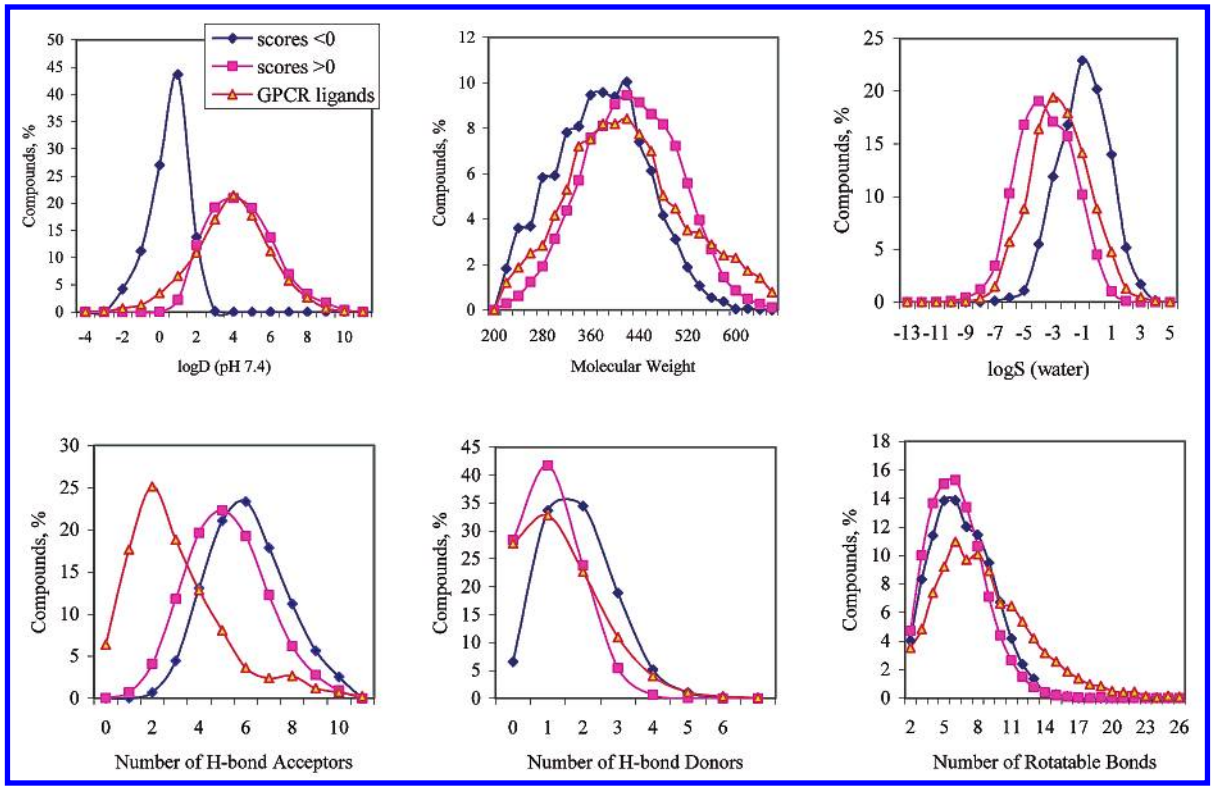

Figure 7. Molecular descriptor distribution profiles for 5736 GPCR-active compounds (orange), compared to 30 000 compounds with neural net positive (pink) and 10 000 negative (blue) scores.

scoring set. This is especially evident for such key parameters as $LogD_{74}$, $LogS_w$, and the number of hydrogen-bond donors. Interestingly, the shifts in distribution of negatively scoring compounds relative to positively scoring compounds resemble the shifts in mean values of the parameters calculated, respectively, for GPCR(−) and GPCR(+) compounds (see Table 4).

We have also assessed a comparative predictive power of the simple filtering standard procedure and the neural network selection algorithm developed in this work. The simple filtering procedure is usually performed by selecting compounds within the assigned boundary values of a specific parameter. In Figure 5, we show an example of such filtering using $logD_{74}$ as the selection parameter. Three database sets, known GPCR-active ligands (orange), compounds with positive neural net scores (pink), and compounds with negative neural net scores (blue), were analyzed. The upper threshold boundary (V_{max}) does not affect the quality of the

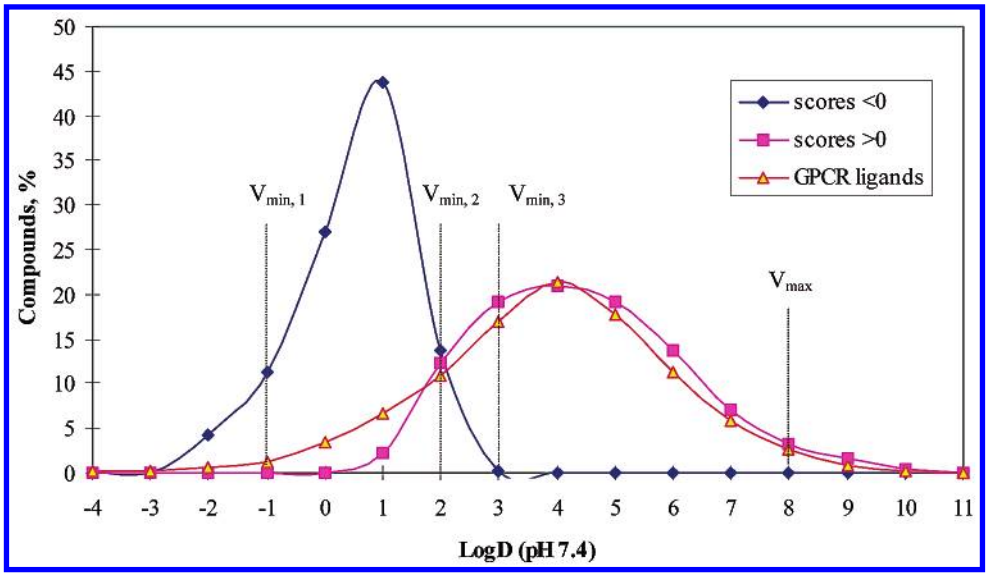

Figure 8. Difference between the simple filtering and the neural net classification procedures. $LogD_{74}$ distribution profiles for 5736 GPCR-active compounds (orange), compared to 30 000 compounds with positive scores (pink) and 10 000 compounds with negative scores (blue).

active/nonactive separation, and its position can be used to exclude compounds with extreme values of a parameter. A critical issue is to determine the optimal position of the lower limit (V_{min}). If the lower selection boundary is set to the $LogD_{74}$ value of −1, to maximally include active compounds, then approximately 30−35% of GPCR nonactives (see Figure 1) and over 80% of all compounds with negative neural net scoring will also be included into this category (Figure 8). Setting a most stringent threshold, $V_{min,3}$, to $LogD_{74} = +3$, to maximally exclude nonactive compounds (Figure 1), results in the exclusion of approximately 30−40% of both the active and positive scoring compounds (Figure 8). Finally, even the most optimal threshold position ($V_{min,2}$) will lead to the inclusion of about 10% of the nonactive and negatively scored compounds and to the loss of the same percentage of the active compounds. Analogous simple discrimination analyses with other descriptors show that the quality of the separation between active and nonactive compounds is even worse and results in a greater error factor. The conclusion of this analysis is that the neural network classification algorithm taking into account the whole assembly of connections between several input variables provides for a more refined selection of compounds with preferable target biased properties.

Lead-Likeness vs Drug-Likeness. Many focused combinatorial library design programs utilize filters based on the well-known Lipinski "rule of five". However, this rule was derived from analyzing drugs, not leads. Having been filtered according to the drug-based empirical rules, the hits obtained in the process of primary HTS, did not prove to be easily amenable for traditional medicinal chemistry optimization.[8] This problem was discussed in a recent paper,[9] where it was suggested that lead-like libraries should be designed with lower MW and lower LogP profiles, as opposed to drug-like libraries. In this work, we used a training set comprising both leads and drugs acting on the receptors of the GPCR family. Therefore, our neural network model generates relationships amenable for selection of lead-like libraries of potential GPCR ligands suitable to further optimization.

CONCLUSIONS

Our observations indicate the presence of a combination of some specific physicochemical features that distinctly differentiate the GPCR-ligands from the compounds belonging to other target-specific classes. Using these findings, we created a neural network classification model with an excellent discrimination power. In a control experiment with a random selection of GPCR(+) and GPCR(−) test compounds, up to 92% of GPCR-ligands and 93% of non-GPCR-ligands were correctly predicted.

The major goal of developing this classification algorithm was to create the effective mechanism for guiding design of combinatorial libraries with preferential GPCR-activity or, more specifically, with a significantly enhanced hit rate. It is anticipated that this general model will be an extremely useful method in constraining the size of combinatorial libraries and in the collection, manipulation, and use of the data that are generated. In general, this neural network based filter is most useful in conjunction with the drug/nondrug filter reported earlier.[1,2] The possibility of library enrichment with potential GPCR biased compounds will help speed up the development of new drugs acting through GPCR therapeutic targets. It should be noted, however, that the score function generated by this particular neural network model does not define the receptor type specificity of the positively identified compounds. A receptor specific model can be developed using neural network training set with the receptor specific compounds.

ACKNOWLEDGMENT

We thank Prof. Hugo Kubinyi (University of Heidelberg, Germany) for invaluable comments and discussion.

REFERENCES AND NOTES

(1) Sadowski, J.; Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.

(2) Ajay, A.; Walters, W. P.; Murcko, M. A. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.

(3) Ajay; Bemis, G. W.; Murcko, M. A. Designing libraries with CNS activity. *J. Med. Chem.* **1999**, *42*, 4942−4951.
(4) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(5) Sternberg, P.; Luthman, K.; Ellens, H.; Lee, C. P.; Smith, Ph. L.; Lago, A.; Elliott, J. D.; Artursson, P. Prediction of the intestinal absorption of endothelin receptor antagonists using three theoretical methods of increasing complexity. *Pharm. Res*. **1999**, *16*, 1520−1526.
(6) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. A knowledge based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *J. Comb. Chem*. **1999**, *1*, 55−68.
(7) Oprea, T. L. Property distribution of drug-related chemical databases. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 251−264.
(8) Teague, S. J.; Davis, A. M.; Leeson, P. D.; Oprea, T. I. The design of leadlike combinatorial libraries. *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 3743−3748.
(9) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is there a difference between leads and drugs? A historical perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308−1315.
(10) van de Waterbeemd, H.; Smith, D. A.; Beaumont, K.; Walker, D. K. Property-based design: optimization of drug absorption and pharmacokinetics. *J. Med. Chem.* **2001**, *44*, 1313−1333.
(11) Gudermann, T.; Nurnberg, B.; Schultz, G.;. G-protein-coupled receptors and G-proteins as primary components of transmembrane signal transduction. Part 1. G-protein-coupled receptors: structure and function. *J. Mol. Med.* **1995**, *73*, 51−63.
(12) Hamm, H. The many faces of G protein signaling. *J. Biol. Chem*. **1998**, *273*, 669−672.
(13) Howard, A. D.; McAllister, G.; Feighner, S. D.; Liu, Q.; Nargund, R. P.; Van der Ploeg, L. H. T.; Patchett, A. A. Orphan G-protein-coupled receptors and natural ligand discovery. *Trends Pharm. Sci.* **2001**, *22*, 132−140.
(14) Ensemble database of biologically active compounds. *Prous Science* **2002**. URL: http://www.prous.com.
(15) Trepalin, S. V.; Yarkov, A. V. CheD: chemical database compilation tool, Internet server, and client for SQL servers. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 100−107.
(16) Trepalin, S. V.; Gerasimenko, V. A.; Kozyukov, A. V.; Savchuk, N. Ph.; Ivaschenko A. A. New diversity calculations algorithms used for compound selection. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 249−258.
(17) Raevsky, O. A.; Trepalin, S. V.; Trepalina, H. P.; Gerasimenko, V. A.; Raevskaya, O. E. SLIPPER-2001 - software for predicting molecular properties on the basis of physicochemical descriptors and structural similarity. *J. Chem. Inf. Comput. Sci.* **2002**, Published on web, January.
(18) Raevsky, O. A. Molecular lipophilicity calculations of chemically heterogeneous chemicals and drugs on the basis of structural similarity and physicochemical parameters. *SAR QSAR Environ. Res*. **2001**, *12*, 367−381.
(19) NeuroDimension, Inc., 2001. URL: http://www.nd.com.
(20) Brown, R. D.; Martin, Y. C. Use of structure−activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 572−584.
(21) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand−receptor binding. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 1−9.
(22) Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of biological activity profiles using substructural analysis and genetic algorithms. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 165−179.

CI025538Y