



Fitting the complexity of GPCRs modulation into simple hypotheses of ligand design

Chiara Custodi^a, Roberto Nuti^a, Tudor I. Oprea^{b,c}, Antonio Macchiariulo^{a,*}

^a Dipartimento di Chimica e Tecnologia del Farmaco, Università di Perugia, via del Liceo 1, 06123 Perugia, Italy

^b Division of Biocomputing, Department of Biochemistry and Molecular Biology, University of New Mexico School of Medicine, Albuquerque, NM, USA

^c Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark

ARTICLE INFO

Article history:

Accepted 2 July 2012

Available online 20 July 2012

Keywords:

Drug design

Chemoinformatics

Structure–activity relationships

G-protein coupled receptors

Drug discovery

ABSTRACT

G-protein coupled receptors (GPCRs) are a large family of membrane-bound receptors that mediate a wide range of physiologic responses to hormones, neurotransmitters and dietary lipids, which represent an important class of drug targets. Significant chemical space regions have been explored both in the academia and by pharmaceutical companies, in the quest for new GPCR modulators as potential therapeutic agents. This accumulated body of evidence provides new opportunities to evaluate potential features of GPCR agonists and antagonists, and how to distinguish them. In this study, the chemical space covered within the WOMBAT database by GPCR modulators was investigated with the aim of identifying specific molecular determinants that distinguish GPCR agonists from antagonists.

While instrumental to get insights into the design strategies of GPCRs modulators, the results of this study provide novel clues on the molecular mechanisms that underlie the complexity of GPCR modulation.

© 2012 Elsevier Inc. All rights reserved.

1. Introduction

G-protein coupled receptors (GPCRs) are membrane proteins involved in the transmission of signaling pathways across cell membranes [1]. Although activated by a diverse set of ligands that include photons (e.g., rhodopsin), ions (e.g., proton or Ca²⁺ sensing GPCRs) [2,3], the vast majority of GPCR ligands are amino acids, fatty acids, steroids and neurotransmitters. GPCRs share a common structural organization consisting of seven-transmembrane (TM) domains, connected by extracellular (ECL) and intracellular (ICL) loops. They are commonly classified into five major families, with family A, also known as Rhodopsin family, being the largest one [4]. What makes GPCRs particularly attractive to pharmaceutical companies as well as academic institutions, is that they are perceived as high druggable targets and less than half of the human GPCRs have been exploited in drug therapy, indicating that there is a potentially significant therapeutic potential to be derived from modulating GPCRs [5–9].

The last decade, in particular, has witnessed outstanding breakthroughs in the understanding of the structural, conformational and mechanistic aspects of GPCRs, with crystallization and biophysical studies providing unprecedented opportunities to boost

GPCRs based drug discovery [10]. Collectively, these studies sustain the conformational complexity of GPCRs, challenging the traditional view of the receptors as bimodal switches of inactive and active states [11]. Thus, according to the paradigm of conformational complexity, GPCR modulation occurs through a continuum of conformational states that feature specific energy landscapes of the receptors [12].

As a consequence, different ligands stabilize diverse active and resting conformations of GPCRs, leading to a differential modulation of signaling across cell membrane. Lefkowitz and coworkers have recently provided experimental support to this concept, evidencing the presence of distinct conformational changes in the β 2-adrenergic GPCR even upon binding of similar ligands [13].

From a pharmacological point of view, agonists are defined as ligands endowed with affinity and positive efficacy at the GPCR, binding to active conformations of the receptor and promoting signal transduction across membrane.

Different agonists may thus bind to and stabilize distinct active conformations of GPCRs, thereby promoting the recruitment of diverse G-protein isoforms for coupling differential signaling pathways. Recent studies suggest that agonists can even activate G-protein independent pathways, introducing the concept of ligand-induced selective signaling (LiSS) as a novel paradigm of GPCR signaling [14–17]. For instance, D2 GPCR agonists have been recently discovered that display selective signaling via the adaptor proteins β -arrestin-2 [18]. These compounds showed

* Corresponding author. Tel.: +39 075 585 5160; fax: +39 075 585 5114.

E-mail address: antonio@chimfarm.unipg.it (A. Macchiariulo).

Table 1

Composition of subclasses of ligands with different pharmacological profile according to the type of GPCR-A. In brackets it is reported the number of ligands for each subclass before balancing the dataset.

Subclass	GPCRs-A	N° agonists	N° antagonists	N° inverse agonists
i	Adrenergic	389 (389)	389 (3031)	0
ii	Dopaminergic	288 (288)	288 (3931)	0
iii	Serotonergic	504 (504)	504 (4841)	64
iv	Melatonergic	93 (93)	93 (882)	0
v	Histaminergic	80 (80)	80 (1040)	8
vi	Muscarinic	150 (150)	150 (1450)	0
vii	Purinergic	327 (327)	327 (3209)	0
viii	Prostanoid	172 (172)	172 (641)	0
ix	Fatty acids (LPA)	22 (22)	22 (118)	0
x	Cannabinoid	75 (75)	75 (792)	6
xi	Peptidergic	383 (383)	383 (6086)	19
xii	Hormone	116 (116)	116 (248)	0

antipsychotic-like activity without motoric side effects in inbred C57BL/6 mice.

Antagonists are defined as compounds endowed with affinity and no efficacy at the GPCR, binding to resting conformations of the receptor and triggering no signal.

Some GPCRs are endowed with basal activity such as serotonin and cannabinoid receptors [19,20], though functionally active cannabinoid-1 receptors have also been found as expressed in intracellular compartments where they respond to anandamide binding and activate NAADP-dependent calcium pathways [21]. The basal or constitutive activity of GPCRs is explained with an inherent dynamicity of the receptor that may adopt more conformational states in the absence of ligands, efficiently coupling with G-protein signaling [22,23]. In this context, ligands binding to orthosteric site may have negative efficacies and, as such, be more properly defined as inverse agonists [24].

Aside from the important implications of these observations in GPCR drug discovery, the above findings provide new opportunities to study how agonists and antagonists work. To this aim, herein we report a study based on the construction of decision trees that, identifying specific molecular properties able to distinguish GPCR agonists from antagonists, provide clues to further the understanding of molecular mechanisms that underlie the complexity of GPCR modulation, and aid the identification of GPCR modulators with specific pharmacological profiles from virtual screening of large collections of compounds. In particular, the space covered within the WOMBAT database [25] is taken as the dataset source to collect agonists and antagonists for the larger annotated receptors belonging to family A GPCRs (Table 1); whereas either 2D or 3D MOE descriptors are used to map the chemical space. The results are discussed in terms of the simplest hypothesis of molecular descriptors able to distinguish agonists from antagonists for each of the selected GPCRs, as well as regardless of the specific family of receptors. Furthermore, case studies from a thorough literature search are reported that support the proposed simplest hypotheses of agonist and antagonist design, providing mechanistic explanations to the selected molecular descriptors.

2. Results and discussion

Decision trees are classifiers that predict group membership of objects by splitting them according to nodes of questions on their features (Fig. 1) [26].

In this study, while objects are the collected 12 subclasses of GPCR-A ligands (2599 agonists vs. 2599 antagonists), features are the four set of 2D and 3D molecular descriptors. A fifth set of descriptors based on fingerprints representation of ligands was also calculated as instrumental to investigate whether agonists and

antagonists were biased to specific scaffolds in the subclasses of GPCR-A ligands.

Accordingly, we applied decision trees to model and interpret agonism and antagonism of GPCR-A ligands at 12 types of receptors (Table 1), assigning each compound to the group of agonists (group 1) or antagonists (group 2) on the basis of one or more selected descriptors. Although additional improvements to the study could be expected from the inclusion of inverse agonists, the poor number of compounds found in WOMBAT with this functional annotation hampered the extension of the analysis to this pharmacological class of ligands (Table 1).

Table 2 shows the results of decision trees generated for the 12 subclasses of GPCR-A ligands on the basis of the five set of descriptors (see Section 4 and supplementary materials). The resulting decision trees are composed of a different number of nodes, ranging from a minimum of one layer with three nodes and one splitting variable (low complexity decision tree) to a maximum of 4 layers with multiple nodes and splitting variables (high complexity decision tree). In the case of peptidergic GPCRs (subclass xi), no results were found for decision trees generated on the connectivity indices as well as when using dipole and molecular shape descriptors. Likewise, using fingerprints representation of ligands, only subclasses i, ii, vi, ix, x and xii yielded decision trees composed of different numbers of nodes.

In order to assess the presence of any relationship between the complexity of decision trees and the number of GPCR-A isoforms targeted by each subclass of ligands, the average sequence identity of human receptor isoforms was calculated within the 12 families of GPCRs. As a result, no linear trend is observed, suggesting that the low or high complexity of decision trees is not related to the high or low number of conserved residues within GPCR-A isoforms. Accordingly, the putative presence of selective ligands at specific receptor isoforms does not affect the overall result in each subclass.

Since the aim of the work was to infer the simplest hypotheses of ligand design from decision trees that could provide rules of thumbs to aid the development of novel GPCR-A agonists and antagonists, next we defined criteria to select the simplest decision tree. The first criterion, in particular, was the selection of decision trees composed of one layer and three nodes (one parent node and two child nodes). From the inspection of Table 2, it is found that only 10 out of 12 subclasses are compliant with this condition, showing decision trees composed of only one layer with three nodes and one splitting variable. One layer decision trees resulting from fingerprints representation of ligands were composed of a number of nodes ranging from three to seven. Accordingly, only in one case (subclass x, cannabinoid GPCRs) agonists and antagonists were clearly classified into two separate classes of objects, with these latter being biased to specific chemical scaffolds (Table s11, supplementary materials).

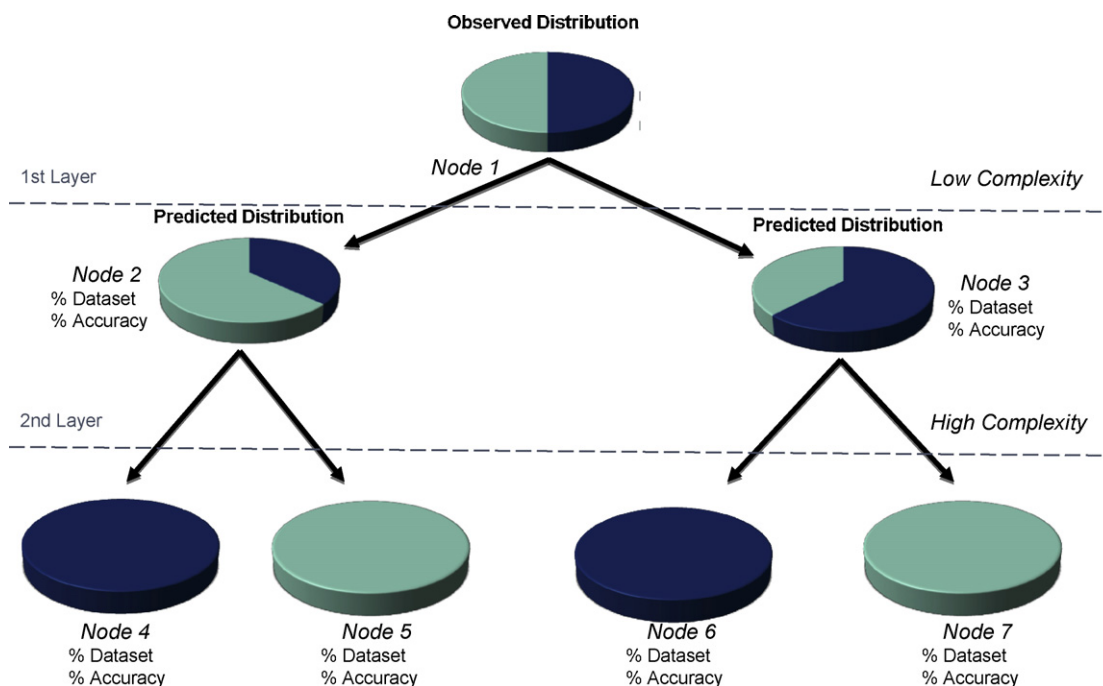


Fig. 1. General scheme of a decision tree. The number of correct classifications in a node gives the percentage (%) of accuracy; the percentage (%) of the dataset gives the number of compounds classified in a child node from a parent node.

As a second criterion, we selected decision trees with accuracy (% of correct classifications) of agonists and antagonists in the node above 80%. Although the value of this threshold is arbitrary, we thought that lower values would have increased the complexity of the hypothesis of ligand design, reducing its statistical significance and increasing the risk of fallacy of the study.

Five subclasses of GPCR-A ligands were thus discarded, whereas subclasses ii, vi, viii, ix, and x were selected for further investigations, matching the second criteria (Table S5 contains the list of splitting nodes for all subclasses of GPCR-A ligands).

Among these, subclass ii is defined by agonists and antagonists acting at dopaminergic D1-like and D2-like GPCRs. The simplest decision tree that correctly predicts group membership of these ligands is based on the number of rigid bonds (RGB) in the compound (Fig. 2, the mathematical definition of RGB is reported elsewhere) [27]. Hence, as rule of thumb, compounds with $RGB \geq 21$ (Table 3) are mostly antagonists (average accuracy $83.5 \pm 1.7\%$), whereas ligands with $RGB < 21$ are generally agonists (average accuracy $80.3 \pm 0.7\%$).

Since 6 RGB correspond roughly to 1 ring, here we are essentially finding that a number of 3.5 rings in a compound is the cut-off

value between agonists and antagonists. This result is in nice agreement with the proposed binding mode of agonists and antagonists at dopaminergic GPCRs. Indeed, Teeter et al. reported that the binding site of dopaminergic receptors is lined with aromatic residues that show reduced side chain conformational flexibility [28]. As a consequence, they suggested that aromatic residues and different rigid shapes of agonists and antagonists could mutually adjust to determine two diverse binding modes, with the former interacting into the binding site perpendicular to the membrane plane and the latter parallel to the membrane plane. Thus, a reduced conformational flexibility of dopaminergic GPCRs binding sites combines with a different number of rigid bonds (RGB) in agonists and antagonists, eventually causing diverse binding modes and functional outcomes.

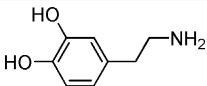
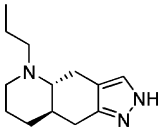
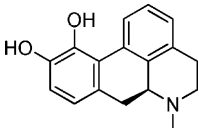
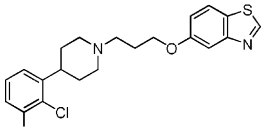
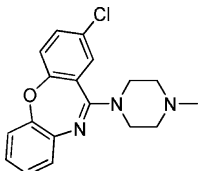
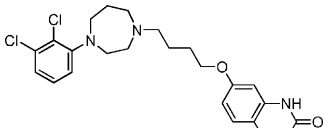
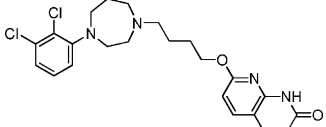
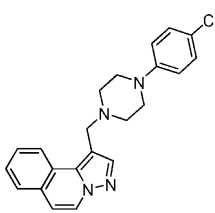
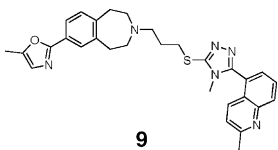
Some D2 receptor ligands act as antagonists of Gi-regulated cAMP production and partial agonists for β -arrestin-2 interactions (4, 6, 7) [18]. While the compliance of these compounds to the rule of thumb of D2 antagonism ($RGB \geq 21$) is in nice agreement with their antagonistic activity at Gi-coupled D2 (Table 3), diverse binding modes of these ligands to D2 receptor may provide an explanation to the β -arrestin-biased activity [13].

Table 2
Average sequence identities of GPCR-A isoforms and number of layers of the decision tree for the 12 subclasses of GPCR-A ligands according to the four set of descriptors. In parenthesis, it is reported the number of nodes of the decision tree.

Subclass	Pharmacophoric topological properties (2D)	Shape connectivity (2D)	Dipole shape (3D)	Surface volume (3D)	MACCS fingerprints	Average sequence identities
i	1 (3)	2 (5)	1 (3)	2 (5)	1 (7)	35.0%
ii	1 (3)	3 (6)	2 (5)	2 (5)	1 (4)	31.6%
iii	4 (9)	2 (5)	2 (5)	2 (5)	–	28.8%
iv	3 (7)	2 (7)	2 (5)	1 (3)	–	46.1%
v	2 (5)	2 (7)	2 (7)	3 (7)	–	25.5%
vi	1 (3)	1 (3)	1 (3)	1 (3)	1 (4)	47.7%
vii	1 (3)	3 (7)	1 (3)	4 (11)	–	26.0%
viii	2 (5)	1 (3)	2 (5)	1 (3)	–	27.9%
ix	1 (3)	2 (5)	2 (5)	1 (3)	1 (4)	32.4%
x	1 (3)	1 (3)	1 (3)	1 (3)	1 (3)	39.1%
xi	2 (7)	–	–	1 (3)	–	24.6%
xii	1 (3)	1 (3)	2 (5)	3 (7)	1 (4)	49.1%

Table 3

Examples of correctly assigned agonists and antagonists from the simplest decision tree of dopaminergic ligands (subclass ii).

Compound	Pharmacological profile	Splitting variable "RGB"
 1 Dopamine	Agonist	6
 2	Agonist	15
 3	Agonist	20
 4 UNC9994[a]	Gi antagonist β -arrestin partial agonist	22
 5	Antagonist	23
 6 UNC0006[a]	Gi antagonist β -arrestin partial agonist	24
 7 UNC9975[a]	Gi antagonist β -arrestin partial agonist	24
 8	Antagonist	27
 9	Antagonist	33

^aCompounds taken from Allen et al. were not included in the training set of the study.

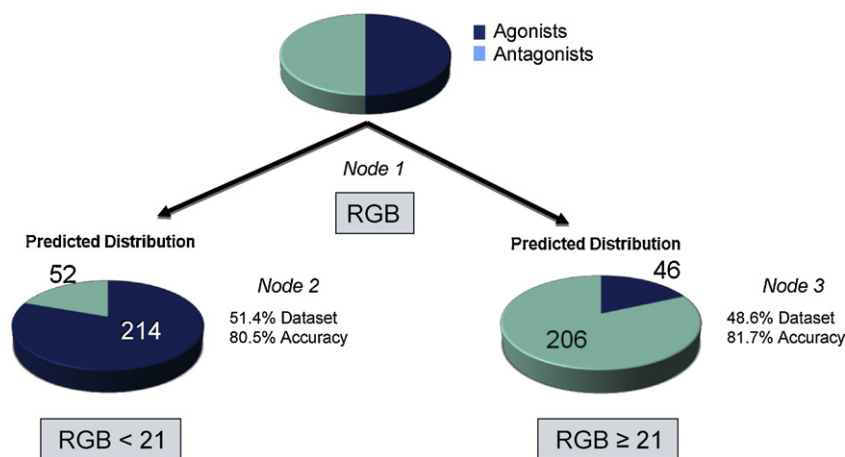


Fig. 2. Decision tree of dopaminergic agonists and antagonists.

Table 4
Examples of correctly assigned agonists and antagonists from the simplest decision tree of muscarinic ligands (subclass vi).

Compound	Pharmacological profile	Splitting variable "a.heavy"	Splitting variable "chi0v.C"	Splitting variable "SASA"	Splitting variable "Vsurf.V"
<chem>CC(=O)OCC[N+](C)(C)C</chem> 10 Acetylcholine	Agonist	10	5.91	373.55	754.62
<chem>Cc1nc2c(c1)ccc2n3c2ccccc3</chem> 11	Agonist	15	7.84	416.36	510.25
<chem>COc1ccc(cc1)CN2CCc3ccccc3CC2</chem> 12	Agonist	22	12.83	590.69	434.50
<chem>C[C@H]1CC[C@@H]2[C@@H]3CC[C@H]4[C@@H]1CC[C@@H]2C(=O)O4</chem> 13	Antagonist	25	15.03	616.46	947.37
<chem>Clc1ccc(cc1)S(=O)(=O)N2CC[C@H]3CC[C@@H]2C(=O)O3</chem> 14	Antagonist	30	15.14	733.72	1008.37
<chem>CCN1CC[C@H]2CC[C@@H]1CC[C@@H]2S(=O)(=O)c1ccc(cc1)S(=O)(=O)c2cc3ccccc3cc2</chem> 15	Antagonist	40	18.39	795.04	1230.00

Muscarinic ligands (subclass vi) are the only subclass where decision trees correctly predict group membership of agonists and antagonists regardless to the set of 2D and 3D descriptors used (Table 4).

Indeed, the selected four descriptors are highly intercorrelated in this subclass of ligands (lowest correlation coefficient observed: $r^2=0.9$) and comprise the number of heavy atoms (a.heavy, set of pharmacophoric and topological properties), carbon valence

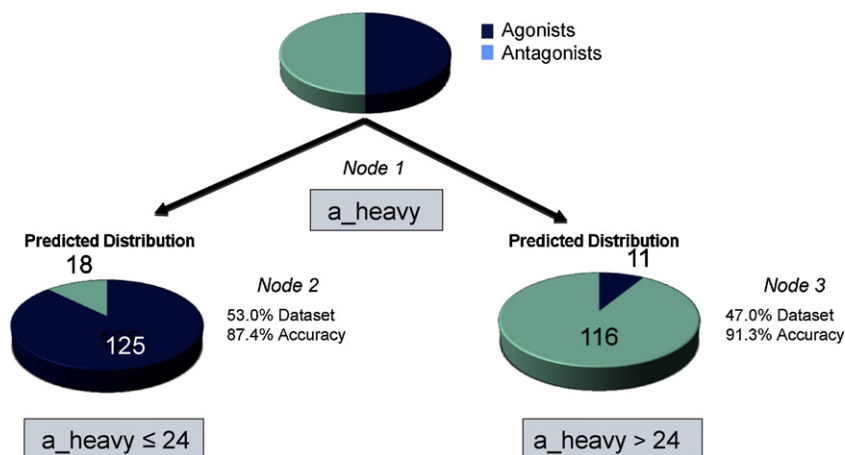
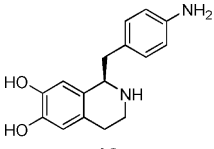
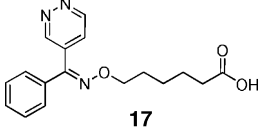
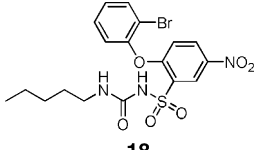
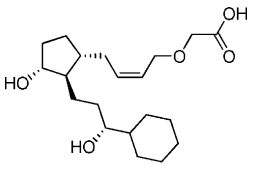
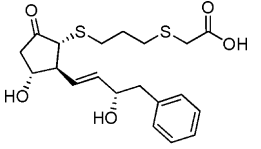
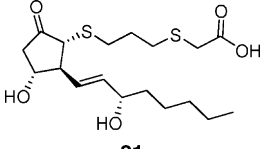


Fig. 3. Decision tree of muscarinic agonists and antagonists according to the number of heavy atoms (a_{heavy}).

Table 5

Examples of correctly assigned agonists and antagonists from the simplest decision tree of prostanoid ligands (subclass viii).

Compound	Pharmacological profile	Splitting variable "KierFlex"	Splitting variable "Vsurf_HB1"
 16	Antagonist	2.68	184.37
 17	Antagonist	5.06	273.12
 18	Antagonist	8.01	228.75
 19	Agonist	8.61	425.37
 20	Agonist	9.30	512.62
 21	Agonist	12.61	539.25

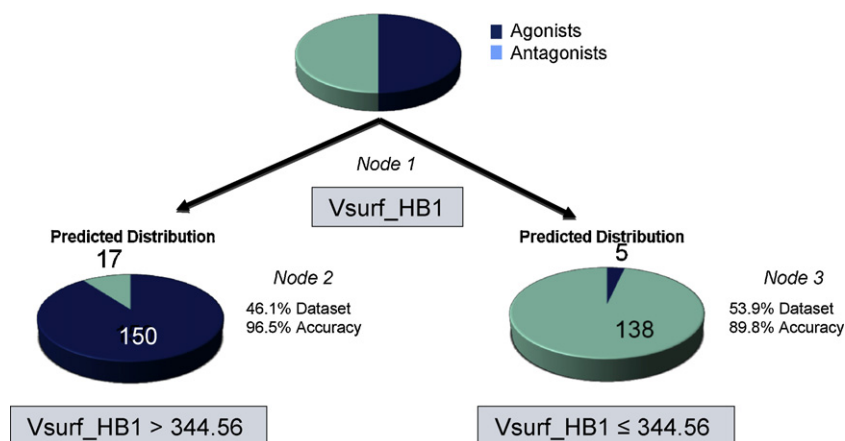


Fig. 4. Decision tree of prostanoid agonists and antagonists according to the hydrogen bond donor capacity of the ligand (Vsulf_HB1).

connectivity index (chi0v.C, set of connectivity indices), solvent accessible surface area (SASA), and the interaction field volume (Vsulf.V, set of volume surface descriptors). All of them are related to the size of ligands, with antagonists showing a number of heavy atoms (a_heavy) > 24 and agonists having a number of heavy atoms (a_heavy) ≤ 24 (Fig. 3).

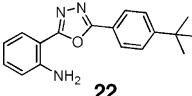
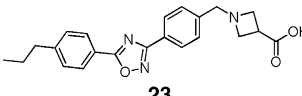
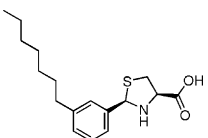
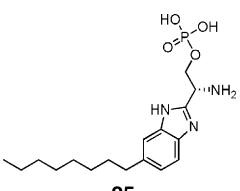
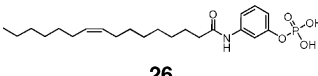
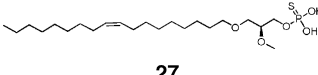
Noteworthy, relying on the size of the modulator, the simplest design hypothesis of muscarinic ligands agrees with two traditional analog-based strategies to design antagonists in medicinal chemistry: (i) design ligands that reach additional binding regions within the orthosteric cleft of the receptor and not used by the endogenous modulator; (ii) design compounds that, by

occupying an additional site close to the orthosteric region, act as a “shield” or umbrella, preventing the endogenous modulator from binding to the GPCR. As a consequence of both strategies, antagonists are very often larger than the endogenous ligands and agonists.

Subclass viii is defined by agonists and antagonists acting at prostanoid GPCRs. Two decision trees are found as being compliant to the criteria of the simplest hypothesis of agonist and antagonist design (Table 5).

The first tree arises from the set of connectivity indices and shows the Kier molecular flexibility index as unique splitting node [29]. The second tree originates from the set of volume surface

Table 6
Examples of correctly assigned agonists and antagonists from the simplest decision tree of lisophosphatidic acid (LPA) GPCRs (subclass ix).

Compound	Pharmacological profile	Splitting variable “b.1rotN”	Splitting variable “Vsulf.ID7”
 22	Antagonist	2	0.28
 23	Antagonist	7	1.28
 24	Antagonist	8	1.39
 25	Agonist	11	3.52
 26	Agonist	16	3.23
 27	Agonist	22	4.19

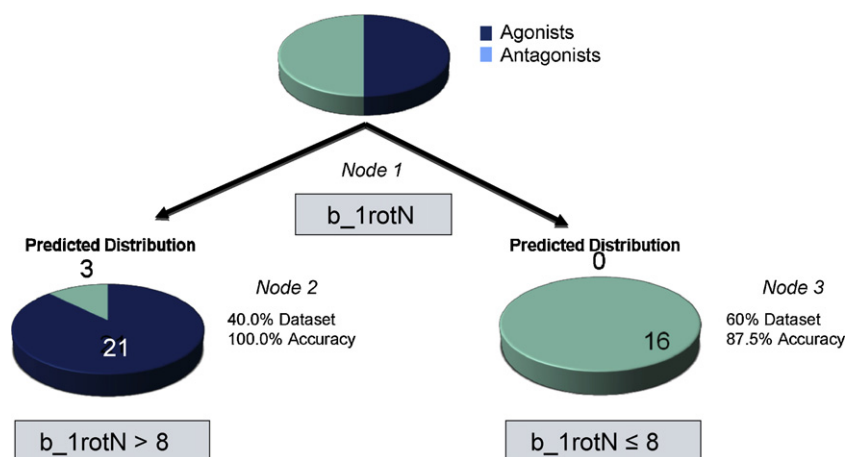


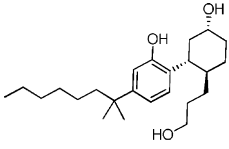
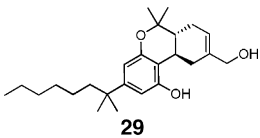
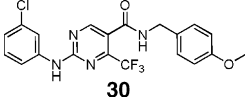
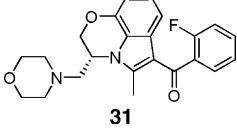
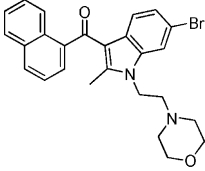
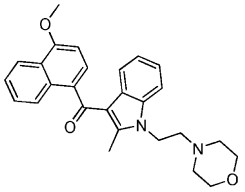
Fig. 5. Decision tree of LPA GPCRs agonists and antagonists according to the number of rotatable single bonds (b_{1rotN}).

descriptors and has the hydrogen bond donor capacity of the ligand (V_{surf_HB1}) as splitting node [30]. The correlation coefficient (r^2) between the two descriptors is 0.7, indicating that they encode similar information. As a rule of thumb, either compounds with

KierFlex higher than 8.09 or V_{surf_HB1} higher than 344.56 \AA^2 act as agonists, the other way round they are antagonists. Although both decision trees are compliant to the criteria of the simplest hypothesis of ligand design, the one generated on V_{surf_HB1} shows better

Table 7

Examples of correctly assigned agonists and antagonists from the simplest decision tree of cannabinoid ligands (subclass x).

Compound	Pharmacological profile	Splitting variable "RGB"	Splitting variable "dipole"
 28	Agonist	12	0.51
 29	Agonist	16	0.94
 30	Agonist	18	1.60
 31	Antagonist	26	3.86
 32	Antagonist	27	5.31
 33	Antagonist	27	6.36

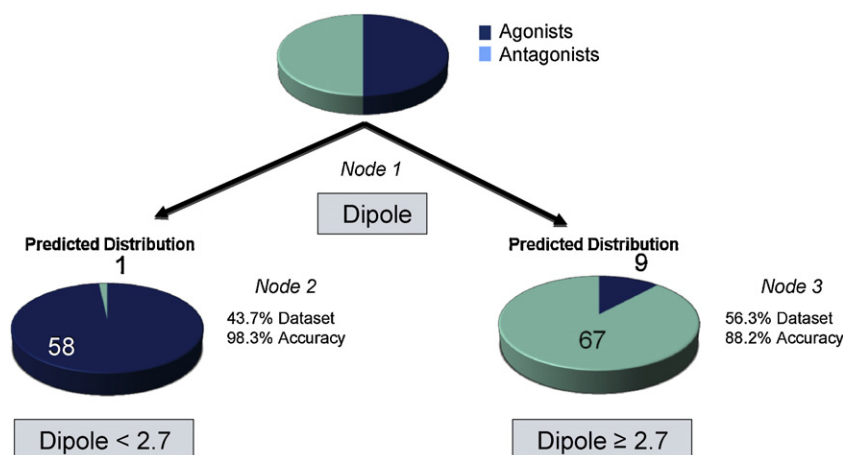


Fig. 6. Decision tree of cannabinoid agonists and antagonists according to the dipole moment (dipole).

performance concerning the accuracy of agonists and antagonists in the node (Fig. 4).

Remarkably, this decision tree is in nice agreement with previous works on prostanoid GPCR agonists evidencing the importance of hydrogen bonding with an arginine residue of the seventh transmembrane helix (Arg309 in the case of prostaglandin EP3 α receptor) for the functional activation of the receptor [31–33].

Lisophosphatidic acid (LPA) GPCRs are the targets of ligands in the subclass ix. Again, two decision trees are compliant to the criteria of the simplest hypothesis of ligand design (Table 6), with one being based on the number of rotatable single bonds (b_{1rotN}) in the compounds and the other on the distance between the molecular center of mass and the barycenter of hydrophobic interaction regions (hydrophobic integrity moment, $Vsurf_ID7$) [30]. As a result, LPA agonists occur with b_{1rotN} higher than 8 or $Vsurf_ID7$ higher than 1.77 Å, vice versa ligands are LPA antagonists (Fig. 5).

The two molecular descriptors are highly correlated ($r^2 = 0.8$), suggesting that for this series they are redundant. These decision trees, however, suffer from the poor number of instances in the subset (22 agonists and 22 antagonists) which limits their statistical significance and applicability. As a consequence, the observed rules seem to apply mostly for lipid LPA agonists and non lipid LPA antagonists, though lipid LPA antagonists have also been reported in literature [34].

Subset x is made of ligands acting at cannabinoid GPCRs. In this case, three decision trees are compliant to our two criteria of the simplest hypothesis of ligand design (Table 7). While the first and second trees are generated on 2D and 3D molecular descriptors, the third decision tree arises from fingerprints representation of ligands. In particular, the first tree is the result of pharmacophoric and topological descriptors displaying the number of rigid bonds (RGB) as splitting variable, the second tree arises from the set of dipole and shape descriptors and has the dipole moment calculated from the partial charges of the molecule (dipole) as splitting variable.

Thus, as a rule of thumb, cannabinoid ligands either with the number of rigid bonds higher than 20 or dipole moment higher than 2.70 Debye act as antagonists, the other way round they are agonists (Fig. 6). Supporting the dipole moment as splitting variable, it has been suggested the presence of a local electrostatic field as characterizing the ligand binding pocket of cannabinoid receptors. Accordingly, different dipole moments on cannabinoid ligands may combine with the local electrostatic field of the binding site, promoting diverse activation states of these GPCRs [35,36].

Beside these characteristics, however, the results of fingerprints representation of cannabinoid ligands pinpoint that the pharmacological profile of agonist and antagonist may be also linked to specific structures, being these biased to most representative chemical scaffolds (Table s11, supplementary materials).

We next wondered whether specific molecular determinants could exist to distinguish agonists from antagonists, regardless to the specific family of GPCRs. Table 8 shows that none of the decision trees generated for the whole dataset satisfies the criterion of having accuracy above 80%, with fingerprints yielding no results at all. These findings evidence the better performance of previous local models over the global model to infer simplest hypotheses of ligand design for GPCR modulators.

Nevertheless, it is of interest the observation that a poor conformational flexibility (high number of rigid bonds, RGB) shows the better accuracy in distinguishing antagonists from GPCRs-A agonists. We have previously discussed this molecular descriptor in the context of dopaminergic agonists and antagonists, suggesting that different conformational flexibilities of ligands may combine with multiple conformational states of the GPCR binding site and causing diverse binding modes and signaling outcomes. In view of the results of the global model, we can tentatively extend this consideration to all of the selected GPCRs-A.

As far as it concerns the performance of the five set of molecular descriptors in distinguishing GPCR agonists from antagonists in local models, it is worth noting that overall 2D descriptors perform

Table 8
Accuracy (average \pm standard deviation percentage of correct classifications after 10-fold cross-validation analysis) of agonists and antagonists in the node according to the global model of decision trees.

GPCRs-A	Accuracy of agonists	Accuracy of antagonists	Splitting variable (agonists)	Splitting variable (antagonists)
Pharmacophoric-topological properties (2D)	62.90 \pm 0.007%	65.91 \pm 0.007%	RGB < 21	RGB \pm 21
Shape-connectivity (2D)	56.34 \pm 0.003%	56.83 \pm 0.003%	Zagreb < 145	Zagreb \geq 145
Dipole-shape (3D)	57.51 \pm 0.003%	61.40 \pm 0.004%	FASA-H < 0.857	FASA-H \geq 0.857
Surface-volume (3D)	54.29 \pm 0.002%	68.93 \pm 0.012%	Vsurf-WP5 \geq 8.563	Vsurf-WP5 < 8.563
MACCS fingerprints	–	–	–	–

Table 9
 Accuracies (average \pm standard deviation percentage of correct classifications after 10-fold cross-validation analysis) of agonists and antagonists in the node of selected decision trees.

Subclass	Pharmacophoric topological properties (2D)	Shape connectivity (2D)	Dipole shape (3D)	Surface volume (3D)	MACCS fingerprints
i	82.9 \pm 0.5% (Ago.) 78.3 \pm 0.7% (Ant.)	-	63.5 \pm 5.8% (Ago.) 81.2 \pm 9.5% (Ant.)	-	-
ii	80.3 \pm 0.7 (Ago.) 83.5 \pm 1.7 (Ant.)	-	-	-	-
iii	-	-	-	-	-
iv	-	-	-	57.1 \pm 11.1% (Ago.) 61.0 \pm 6.4% (Ant.)	-
v	-	-	-	-	-
vi	85.5 \pm 3.7% (Ago.) 91.1 \pm 0.5% (Ant.)	88.9 \pm 0.9% (Ago.) 85.6 \pm 0.2% (Ant.)	80.4 \pm 3.8% (Ago.) 89.0 \pm 12.5% (Ant.)	81.4 \pm 2.1% (Ago.) 92.0 \pm 0.9% (Ant.)	-
vii	90.7 \pm 1.4% (Ago.) 67.7 \pm 0.7% (Ant.)	-	59.7 \pm 3.9% (Ago.) 87.5 \pm 10.9% (Ant.)	-	-
viii	-	83.7 \pm 1.8% (Ago.) 88.9 \pm 2.2% (Ant.)	-	89.3 \pm 0.4% (Ago.) 93.5 \pm 0.6% (Ant.)	-
ix	95.1 \pm 10.4 (Ago.) 85.5 \pm 4.0 (Ant.)	-	-	95.3 \pm 1.7% (Ago.) 90.9 \pm 3.4% (Ant.)	-
x	97.0 \pm 0.9% (Ago.) 84.8 \pm 0.7% (Ant.)	74.6 \pm 2.2% (Ago.) 94.1 \pm 4.3% (Ant.)	98.6 \pm 0.7% (Ago.) 87.4 \pm 1.3% (Ant.)	88.6 \pm 2.3% (Ago.) 77.5 \pm 1.2% (Ant.)	95.0 \pm 6.4% (Ago.) 94.6 \pm 13.3% (Ant.)
xi	-	-	-	57.8 \pm 3.8% (Ago.) 78.9 \pm 10.7% (Ant.)	-

slightly better than 3D descriptors, with the latter being correlated to the former in the four simplest decision trees wherein they are selected as splitting variable (Table 9). This observation, however, may be biased by the crude approximation of calculating 3D descriptors with the global minimum conformation that may not always match the bioactive conformation of ligands (see Section 4 for further details). More in detail, pharmacophoric and topological 2D-properties are selected as splitting variables in four of the simplest decision trees (subclass ii, vi, ix, x), whereas 2D-connectivity indices constitute the best spitting variables in three of the simplest decision trees (subclass vi, viii, x). Interestingly, volume surface 3D-descriptors are found as splitting variables in four of the simplest decision trees (subclass vi, viii, ix, x), suggesting a good performance of this set of descriptors with respect to molecular shape 3D-descriptors.

Regarding the set of fingerprints descriptors, on the one side the poor results obtained with this type of representation of ligands provide evidence that, with the exception of cannabinoid ligands (subclass x), agonists and antagonists are not biased to specific chemical scaffolds in the subclasses of GPCRs-A ligands. On the other side, they show that agonist or antagonist profile is more linked to specific chemical characteristics (encoded by 2D and/or 3D molecular descriptors) rather than to specific chemical structures (encoded by fingerprints).

Ultimately, caveats of this study should also be mentioned. We have already mentioned the emerging complexity of GPCR modulation. Many of the early reports on GPCR ligands lack therefore complete annotations which may result in incomplete or wrong functional evaluation among agonists and antagonists, which may increase the difficulty of rationalization. For instance, common errors of functional annotations in the field of GPCR modulators may concern some agonists that activate G-protein independent pathways or some antagonists that in physiological conditions rather act as inverse agonists. Although we may not rule out that these issues can also affect the 5 selected subclasses of GPCR-A ligands (ii, vi, viii, ix, x), they may certainly in part explain the high complexity and/or low accuracy of decision trees for the remaining 7 subclasses (i, iii, iv, v, vii, xi, xii). Furthermore, we may not exclude that one simple hypothesis of ligand design could not exist for these latter subclasses, but rather a combination of different structural, geometrical, and electronic aspects would affect the functional profile of ligands at these GPCRs-A. Alternatively, it is also possible that the number and/or types of descriptors used in this study do not cover thoroughly the molecular properties of compounds, providing an incomplete map of the chemical space where regions are missing in which GPCR-A agonists and antagonists fall much further apart.

3. Conclusions

In this study, we have examined whether simplest hypotheses of agonist and antagonist design could exist for GPCR-A ligands. Despite the aforementioned caveats of the study, our results have shown that it is possible to identify such hypotheses for discriminating between agonists and antagonists, albeit for selected subclasses of GPCR-A ligands including dopaminergic GPCRs (subclass ii), muscarinic GPCRs (subclass vi), prostanoid GPCRs (subclass ix), LPA GPCRs (subclass ix) and cannabinoid GPCRs (subclass x). Remarkably, case studies from literature support our observations, providing mechanistic explanations to the molecular descriptors that grounded the simplest hypotheses of ligand design.

We have also shown that local models provide better results than a global model of classification to infer simplest hypotheses of ligand design for GPCR modulators, though an interesting

observation could still be made on the general poor conformational flexibility shown by antagonists with respect to agonists.

Although the results of this study may be used as rules of thumbs to aid the identification of novel ligands of these GPCRs-A with specific pharmacological profiles, the future of this work grounds in exploring different possibilities for describing both the chemical space and the functional profile of ligands. For instance, alternative ways to map compounds in the chemical space might lead to a better understanding of their functional profile. Likewise, additional improvements to the study may come from extending the functional classes of ligands to include also inverse agonists, subtype selective GPCR modulators, and selective GPCR signaling ligands, provided that these classes of ligands will become increasingly annotated with the advancement of thorough biological and pharmacological assays. This would allow the broadest possible coverage of functional profiles of GPCRs ligands, eventually dissecting the complexity of GPCR modulation and providing novel clues on the molecular mechanisms that underlie GPCR functioning.

4. Computational methods

4.1. Dataset collection

GPCR ligands were retrieved from “WOMBAT 2008 v11” (*World of Molecular BioActivity*) database, and included 220,733 molecules along with biological activity annotations [25]. In the present study, only family A of GPCRs was considered since members of this class of receptors are the most representative drug targets in the human body [8]. In order to collect ligands with high free energy of binding, we selected only molecules with a ligand efficiency (LE) higher than 3. As a consequence, starting from 220,733 GPCR modulator entries, a total of 39,493 GPCR-A ligands was eventually collected and divided into classes of modulators according to the pharmacological profile of agonism, antagonism and inverse agonism. Further 12 subclasses of modulators were then defined on the basis of the GPCR-A target. Since the number of inverse agonists was poor compared to agonists and antagonists, this class of modulators was not further considered in the analysis. Then, in order to have a balanced dataset of agonists and antagonists in each of the 12 subclasses, an identical number of agonists and antagonists was randomly selected, yielding a total of 2599 agonists and 2599 antagonists (Table 1).

4.2. Ligand preparation

For each molecule of the dataset, we generated a 3D structure using LigPrep 2.0 program as implemented in Maestro [33]. The geometries of all the compounds were energy refined using the OPLS 2005 force field. Each compound was then submitted to a conformational analysis using ConfGen 2.2. During the conformational analysis, the solvent was implicitly treated by considering a dielectric constant of 4. A post-minimization step was also carried out using the Truncated Newton Conjugate Gradient in order to efficiently find conformations with low gradient. The number of iterations was set to 100, while the convergence threshold was set to 0.001.

Since a careful assessment of the bioactive conformation for each of the compound of the dataset is unfeasible given the high amount of ligands, we consider only the global minimum conformer for each compound. Accordingly, only the global minimum conformation for each agonist and antagonist was stored in the dataset and used, in particular, for the calculation of 3D descriptors.

4.3. Molecular descriptors

For the purposes of the study, we selected and calculated 2D, 3D descriptors and fingerprints using MOE v2009-10 (see supplementary materials, Tables S1–S4) [33]. 2D descriptors included two set of descriptors: pharmacophoric and topological properties (44 descriptors); connectivity indices (16 descriptors). 3D descriptors comprised two additional set of descriptors: dipole and molecular shape descriptors (16 descriptors); volume surface descriptors (85 descriptors). Fingerprints were 166 MACCS structural keys that were produced by MOE as sets of integers. Each fingerprint was a bit vector of feature bits 6 words long.

The five set of descriptors were not merged in the study, but they were used to generate four decision trees for each subset of GPCR modulators. The idea here was to assess the overall performance of these classes of descriptors in distinguishing GPCR agonists from antagonists, being aware of the crude approximation introduced in the calculation of 3D descriptors using the global minimum conformation of ligands that may not represent the bioactive conformation. At this regard, however, it should be mentioned that having selected compounds with high ligand efficiency ($LE > 3$), this approximation may be softened by the consideration that most of the ligands may have bioactive conformations very close to the relative global minimum. Indeed compounds with high LE [$LE = (\text{activity})/(\text{molecular weight})$] include moderately active compounds endowed with smaller size (lower molecular weight), and/or highly active compounds with larger size. While a narrow conformational profile may be expected in the former compounds with few conformations around the global minimum, a wider conformational profile is likely to occur in the latter ligands. However, these latter compounds may still have bioactive conformations close to the relative global minimum in order to favor the free energy of binding to the GPCR, thereby accounting for their high activity.

4.4. Decision tree

Decision tree or recursive partitioning is a statistical method that predicts group membership for compounds in a given dataset [37]. The core of decision tree is the partitioning of data into nodes along branches. This is achieved by analyzing the dataset of compounds for which the target property (in our case agonism vs. antagonism) has been used to define the group membership. Ideally, a single node would perfectly divide the dataset according to the group membership of the compounds. If this is not the case, then a set of layers of nodes is constructed in order to separate the dataset in the cleanest way. The degree of accuracy in a node is given by the percentage of correctly classified compounds.

In this study, decision trees were generated using the CHAID algorithm [34] as implemented in the XLSTAT software. This approach, in particular, proceeds in splitting the parent node that contains all the compounds in child nodes which are then split recursively until one of the stopping criteria is met. These include the reaching of a maximum number of five layers of nodes in the tree, the presence of at least one child node containing less than half of the compounds in the parent node, the presence of nodes containing only compounds of one category. For each node, the best splitting variable is selected as the one for which the p -value is lower than a defined threshold of statistical significance set to 0.05.

Once the decision trees were constructed, cross-validation test were run using a leave- n -out approach to assess the statistical significance of the model. Accordingly, for each decision tree, a 10-fold cross-validation analysis was carried out using a validation set containing 10% of the molecules randomly chosen from the original set of compounds. The resulting 48 models were then analyzed

to identify the simplest decision trees, as discussed in the results and discussion section. Confusion matrices of the simplest decision trees (subclass ii, vi, viii, ix, and x; [Tables s6–s10](#)) are reported in the supplementary material.

4.5. GPCR-A sequence similarity

Average sequence identities among different isoforms of selected human GPCR-A subtypes ([Table 1](#)) were assessed using clustal-W2 algorithm and blosum-62 sequence similarity matrix as implemented at the EMBL-EBI server [34].

Acknowledgments

This work was supported in part by NIH grants GM-095952 and MH-084690 (TIO), and by the Villum Foundation CDSB (TIO).

Appendix A. Supplementary data

Supporting information for this article is available. [Tables s1–s4](#) contain the list of molecular descriptors used in this study. [Table s5](#) contains the list of splitting variables for subclasses endowed with one layer of decision tree, but not compliant to the threshold of 80% of accuracy. [Tables s6–s10](#) contain confusion matrices of selected decision trees. [Table s11](#) contains the most representative chemical scaffolds in the agonist and antagonist class of cannabinoid ligands.

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.jmglm.2012.07.002>.

References

- [1] D.M. Rosenbaum, S.G. Rasmussen, B.K. Kobilka, The structure and function of G-protein-coupled receptors, *Nature* 459 (2009) 356–363.
- [2] K. Seuwen, M.G. Ludwig, R.M. Wolf, Receptors for protons or lipid messengers or both? *Journal of Receptor and Signal Transduction Research* 26 (2006) 599–610.
- [3] A.D. Conigrave, H.C. Mun, S.C. Brennan, Physiological significance of L-amino acid sensing by extracellular Ca(2+)-sensing receptors, *Biochemical Society Transactions* 35 (2007) 1195–1198.
- [4] R. Fredriksson, M.C. Lagerstrom, L.G. Lundin, H.B. Schioth, The G-protein-coupled receptors in the human genome form five main families: phylogenetic analysis, paralogon groups, and fingerprints, *Molecular Pharmacology* 63 (2003) 1256–1272.
- [5] C.R. Groom, A.L. Hopkins, The druggable genome, *Nature Reviews Drug Discovery* 1 (2002) 727–730.
- [6] J.P. Overington, B. Al-Lazikani, A.L. Hopkins, How many drug targets are there? *Nature Reviews Drug Discovery* 5 (2006) 993–996.
- [7] P. Imming, C. Sinning, A. Meyer, Drugs, their targets and the nature and number of drug targets, *Nature Reviews Drug Discovery* 5 (2006) 821–834.
- [8] M. Rask-Andersen, M.S. Almen, H.B. Schioth, Trends in the exploitation of novel drug targets, *Nature Reviews Drug Discovery* 10 (2011) 579–590.
- [9] Y. Landry, N. Niederhoffer, E. Sick, J.P. Gies, Heptahelical and other G-protein-coupled receptors (GPCRs) signaling, *Current Medicinal Chemistry* 13 (2006) 51–63.
- [10] R.P. Millar, C.L. Newton, The year in G protein-coupled receptor research, *Molecular Endocrinology* 24 (2010) 261–274.
- [11] B.K. Kobilka, X. Deupi, Conformational complexity of G-protein-coupled receptors, *Trends in Pharmacological Sciences* 28 (2007) 397–406.
- [12] X. Deupi, B.K. Kobilka, Energy landscapes as a tool to integrate GPCR structure, dynamics, and function, *Physiology (Bethesda)* 25 (2010) 293–303.
- [13] A.W. Kahsai, K. Xiao, S. Rajagopal, S. Ahn, A.K. Shukla, J. Sun, et al., Multiple ligand-specific conformations of the beta2-adrenergic receptor, *Nature Chemical Biology* 7 (2011) 692–700.
- [14] T. Kenakin, Agonist-receptor efficacy. II. Agonist trafficking of receptor signals, *Trends in Pharmacological Sciences* 16 (1995) 232–238.
- [15] R.P. Millar, A.J. Pawson, Outside-in and inside-out signaling: the new concept that selectivity of ligand binding at the gonadotropin-releasing hormone receptor is modulated by the intracellular environment, *Endocrinology* 145 (2004) 3590–3593.
- [16] L.R. Rajagopal, K. H.A. Rockman, When 7 transmembrane receptors are not G protein-coupled receptors, *Journal of Clinical Investigation* 115 (2005) 2971–2974.
- [17] S. Ferre, R. Baler, M. Bouvier, M.G. Caron, L.A. Devi, T. Durrour, et al., Building a new conceptual framework for receptor heteromers, *Nature Chemical Biology* 5 (2009) 131–134.
- [18] J.A. Allen, J.M. Yost, V. Setola, X. Chen, M.F. Sassano, M. Chen, et al., Discovery of beta-arrestin-biased dopamine D2 ligands for probing signal transduction pathways essential for antipsychotic efficacy, *Proceedings of the National Academy of Sciences of the United States of America* 108 (2011) 18488–18493.
- [19] V.J. Aloyo, K.A. Berg, W.P. Clarke, U. Spampinato, J.A. Harvey, Inverse agonism at serotonin and cannabinoid receptors, *Progress in Molecular Biology and Translational Science* 91 (2010) 1–40.
- [20] R.G. Pertwee, Inverse agonism and neutral antagonism at cannabinoid CB1 receptors, *Life Sciences* 76 (2005) 1307–1324.
- [21] G.C. Brailoiu, T.I. Oprea, P. Zhao, M.E. Abood, E. Brailoiu, Intracellular cannabinoid type 1 (CB1) receptors are activated by anandamide, *Journal of Biological Chemistry* 286 (2011) 29166–29174.
- [22] Y.X. Tao, Constitutive activation of G protein-coupled receptors and diseases: insights into mechanisms of activation and therapeutics, *Pharmacology & Therapeutics* 120 (2008) 129–148.
- [23] P.L. Prather, Inverse agonists: tools to reveal ligand-specific conformations of G protein-coupled receptors, *Sci STKE* 2004 (2004), pe1.
- [24] T. Kenakin, Efficacy as a vector: the relative prevalence and paucity of inverse agonism, *Molecular Pharmacology* 65 (2004) 2–11.
- [25] R.R.M. Olah, L. Ostropovici, A. Bora, N. Hadaruga, D. Hadaruga, R. Moldovan, A. Fulias, M. Mracec, T.I. Oprea, WOMBAT and WOMBAT-PK: bioactivity databases for lead and drug discovery, in: S.L. Schreiber, T.M. Kapoor, G. Wess (Eds.), *Chemical Biology: From Small Molecules to Systems Biology*, Wiley-VCH, New York, 2007, pp. 760–786.
- [26] C. Kingsford, S.L. Salzberg, What are decision trees? *Nature Biotechnology* 26 (2008) 1011–1013.
- [27] T.I. Oprea, Property distribution of drug-related chemical databases, *Journal of Computer-Aided Molecular Design* 14 (2000) 251–264.
- [28] M.M. Teeter, M. Froimowitz, B. Stec, C.J. DuRand, Homology modeling of the dopamine D2 receptor and its testing by docking of agonists and tricyclic antagonists, *Journal of Medicinal Chemistry* 37 (1994) 2874–2888.
- [29] L.B. Kier, in: B. Boyd, K. Lipkowitz (Eds.), *The Molecular Connectivity Chi Indices and Kappa Shape Indices in Structure–Property Modeling*, D, 1991.
- [30] G. Cruciani, P. Crivori, P.-A. Carrupt, B. Testa, Molecular fields in quantitative structure–permeation relationships: the VolSurf approach, *Journal of Molecular Structure: Theochem* 503 (2000) 17–30.
- [31] C.S. Chang, M. Negishi, N. Nishigaki, A. Ichikawa, Functional interaction of the carboxylic acid group of agonists and the arginine residue of the seventh transmembrane domain of prostaglandin E receptor EP3 subtype, *Biochemical Journal* 322 (1997) 597–601.
- [32] C.S. Chang, M. Negishi, N. Nishigaki, T. Ichikawa, Characterization of functional interaction of carboxylic acid group of agonists and arginine of the seventh transmembrane domains of four prostaglandin E receptor subtypes, *Prostaglandins* 54 (1997) 437–446.
- [33] S. Narumiya, Y. Sugimoto, F. Ushikubi, Prostanoid receptors: structures, properties, and functions, *Physiological Reviews* 79 (1999) 1193–1226.
- [34] G. Tigyi, Aiming drug discovery at lysophosphatidic acid targets, *British Journal of Pharmacology* 161 (2010) 241–270.
- [35] A.M. Ferreira, M. Krishnamurthy, B.M. Moore 2nd, D. Finkelstein, D. Bashford, Quantitative structure–activity relationship (QSAR) for a series of novel cannabinoid derivatives using descriptors derived from semi-empirical quantum-chemical calculations, *Bioorganic and Medicinal Chemistry* 17 (2009) 2598–2606.
- [36] P.G. Willis, O.A. Pavlova, S.I. Chefer, D.B. Vaupel, A.G. Mukhin, A.G. Horti, Synthesis and structure–activity relationship of a novel series of aminoalkylindoles with potential for imaging the neuronal cannabinoid receptor by positron emission tomography, *Journal of Medicinal Chemistry* 48 (2005) 5813–5822.
- [37] Schrodinger LCC Maestro, 2009. New York, NY.