J Comput Aided Mol Des (2015) 29:59–66
DOI 10.1007/s10822-014-9804-5

A desirability function-based scoring scheme for selecting fragment-like class A aminergic GPCR ligands

Ádám A. Kelemen · György G. Ferenczy · György M. Keserű

Received: 7 July 2014 / Accepted: 10 October 2014 / Published online: 19 October 2014

Abstract A physicochemical property-based desirability scoring scheme for fragment-based drug discovery was developed for class A aminergic GPCR targeted fragment libraries. Physicochemical property distributions of known aminergic GPCR-active fragments from the ChEMBL database were examined and used for a desirability function-based score. Property-distributions such as log D (at pH 7.4), PSA, pK_a (strongest basic center), number of nitrogen atoms, number of oxygen atoms, and the number of rotatable bonds were combined into a desirability score (FrAGS). The validation of the scoring scheme was carried out using both public and proprietary experimental screening data. The scoring scheme is suitable for the design of aminergic GPCR targeted fragment libraries and might be useful for preprocessing fragments before structure based virtual or wet screening.

Keywords G-protein coupled receptors · Fragment-based drug discovery · Desirability function · Aminergic fragment library

Abbreviations

FrAGS	Fragment Aminergic GPCR Score
GPCR	G-protein coupled receptor
PSA	Polar surface area
FBDD	Fragment-based drug discovery
HTS	High-throughput screening
FS	Fragment-screening
7TM	Seven-transmembrane
SILE	Size-independent ligand-efficiency
SMILES	Simplified molecular-input line-entry system
EF	Enrichment factor
TPR	True positive rate
TNR	True negative rate
FPR	False positive rate
FNR	False negative rate
ROC	Receiver operating characteristic
$TAAR_1$	Trace-amine receptor subtype 1
$5HT_1$	5-Hydroxy-tryptamine receptor subtype 1

Introduction

Fragment based drug discovery (FBDD) implies the screening of small-sized, simple, polar molecules in high-concentrations. Fragment screening (FS) is an alternative lead-discovery approach to high-throughput screening campaigns with remarkable advantages. Fragment based lead discovery samples the notably smaller fragment space with respect to the lead-like or drug-like space. An illustration of this point is that the number of compounds with up to 17 heavy atoms [1] is about 10^{11} while the estimated number of drug like molecules up to 30 heavy atoms [2] is 10^{60}. Another inevitable aspect is the documented increase of molecular weight and lipophilic character during lead-optimization [3, 4], and fragments as typically soluble, low molecular weight and polar compounds provide more operational freedom in optimizing them to compounds with favorable pharmacokinetic properties.

G-protein coupled receptors (GPCRs) are key modulators of cell-signaling and constitute the largest cell-

Electronic supplementary material The online version of this article (doi:10.1007/s10822-014-9804-5) contains supplementary material, which is available to authorized users.

Á. A. Kelemen · G. G. Ferenczy · G. M. Keserű (✉)
Medicinal Chemistry Research Group, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Magyar Tudósok Körútja 2, Budapest 1117, Hungary
e-mail: keseru.gyorgy@ttk.mta.hu

membrane receptor superfamily. Ligands of these receptors are neurotransmitters, peptides, hormones, etc. GPCRs are either named as seven-transmembrane receptors (7TM) for having a common structure of seven α helices ending in an extracellular N terminus and an intracellular C terminus. The superfamily is classified into A, B, C, and frizzled-type receptors, hereby class A constitutes the largest family, including aminergic, chemokine, glycoprotein hormone receptors and neuropeptide receptors. G-protein coupled receptors are appearing prevalently in drug discovery and representing the therapeutic target for a wide range of small-molecule drugs of central-nervous system, cardiovascular and inflammation related diseases [5]. Aminergic receptors, the largest group of class A, are muscarinic acetylcholine, adrenoceptors, dopamine-, histamine-, serotonin-, octopamine-, and trace amine-receptors. In the present study we collected fragment-like aminergic GPCR ligands from public databases and analyzed their characteristic physicochemical features to derive a scoring scheme suitable to screen compound libraries for aminergic class A GPCR ligands.

Compiling libraries

Active set

GPCR-SARfari [6] was downloaded from the ChEMBL database of the European Bioinformatics Institute [as part of European Molecular Biology Laboratory (EMBL)], and was processed using Knime.com AG's Konstanz Information Miner (Knime) [7]. GPCR SARfari (version 3.00, June 2012) contains 947,914 entries, including molecular structures, and in vitro activity data for GPCR-targets. Collected data were processed in several steps. Only binding data were kept, while data of functional assays and those related to ADME properties were discarded. Activities given as IC_{50}, log IC_{50}, K_i and log K_i were converted to pActivity and were treated on an equal footing with pK_i values. When ligand activities for several species were available only the highest activity was kept. Counter ions of salts were stripped [14], and only one entry was kept with the highest pK_i value. This process resulted in 166,699 entries, containing structurally unique compounds with activity data on several GPCR targets. Size-independent ligand efficiency (SILE) [8] metrics were used to classify compounds as active or inactive on a target. SILE is an empirically derived measure suitable for the comparison of compounds with different sizes:

$$SILE = pK_i/N_{heavy}^{0.3}$$

The cut-off for sorting of GPCR-active fragments was determined by evaluating the relationship between pK_i and SILE in the $8 \leq N_{heavy} \leq 22$ range (see Table 1 in the Supporting Information). SILE $\geq$ 1.95 was accepted as the activity threshold that corresponds to K_i values suitable for fragment-screening activity on GPCR targets (100–10 μM). After removing activities under SILE < 1.95, the resulting data-set of 144,759 entries was used for identifying GPCR-active fragments. Unlike several prevalent methodologies using molecular weight as size-determining definition, our measure was the number of heavy atoms [9]. Molecular-weight based filters penalize halogen and sulphur atoms including sulphonamide-type molecules, due to their proportionately higher atom-weight. Since one heavy atom stands for 13.286 Da [10] molecular weight, the commonly used RO3 [11] cut-off (300 Da) may be converted to $N_{heavy} \leq 22$. Removing molecules with less than 8 or more than 22 heavy atoms resulted in 10,477 unique fragments with listed binding activity data on several GPCRs. Compounds considered to be GPCR-like fragments (2,370), were defined having SILE of at least 1.95 on at least four different GPCR-targets. The active set contains activity data related to 139 different receptors (and subtypes), out of which 87 receptors are aminergic GPCRs [12] (serotonin, β-adrenerg, dopamine, histamine, trace-amine, octopamine, muscarinic-acetylcholine receptors) and 52 other not-aminergic class-A GPCRs (opioid-, adenosine-, sphingosine-1-phosphate-, thromboxane-, melatonin-, orphan-, and angiotensin-receptors). It was found that the majority of the fragments (2,183 out of 2,370, 92 %) had exclusively aminergic GPCR activity data, serving as training set.

Another definition for GPCR-active fragments was also used in order to check how much the property distributions depend on the definition of actives. The activity threshold of SILE $\geq$ 1.95 was replaced by $pK_i = 4$. Note that while pK_i is independent from molecular size, the SILE definition requires higher affinity for larger ligands in accordance with their higher available maximal affinity [8]. Activities over $pK_i \geq 4$ were kept resulting a set of 165,459 entries. Keeping molecules with the number of heavy atoms between 8 and 22 resulted in 10,950 unique fragments. Compounds considered to be GPCR-like fragments, were defined having a pK_i value of at least 4 on at least four different GPCR-targets. This set of 2,478 fragments was used for checking whether the property distributions of the active set are dependent on the method for defining actives (see later).

Reference set

For the determination of the essential GPCR fragment parameters we created a reference set derived from the ChEMBL [13]. Compounds of the active GPCR set were removed from the 1,292,344 ChEMBL (version 16) entries. Ligand structures in salt forms were stripped [14], and only one entry was kept in case of duplicates, resulting in structurally unique compounds. After filtering by

Table 1 Mean, median, and standard deviation values of the calculated physicochemical properties

	Median	Mean	SD
Active set (SILE $\geq$ 1.95)			
PSA (at pH 7.4)	37.30	39.67	21.04
log D (at pH 7.4)	0.97	0.90	1.62
Number of nitrogen atoms	2.00	2.28	1.26
Number of oxygen atoms	1.00	1.08	0.95
Number of rotatable bonds	2.00	2.64	1.89
pK_a (strongest basic)	8.55	7.92	2.85
log P	2.18	2.04	1.53
Hydrogen bond donor count	1.00	0.89	0.96
Hydrogen bond acceptor count	3.00	2.70	1.13
Reference set			
PSA (at pH 7.4)	56.49	60.49	29.28
log D (at pH 7.4)	2.03	1.65	2.24
Number of nitrogen atoms	2.00	2.17	1.48
Number of oxygen atoms	2.00	2.14	1.52
Number of rotatable bonds	3.00	3.36	2.10
pK_a (strongest basic)	7.00	5.97	2.60
log P	2.41	2.20	1.80
Hydrogen bond donor count	1.00	1.27	1.12
Hydrogen bond acceptor count	3.00	3.23	1.50

N_{heavy} = [8; 22], 309,962 fragments remained, out of which 5,000 fragments were picked out randomly to generate a set proportional in size with the active set.

Selection of relevant physicochemical-descriptors and scoring

The examination of molecular physicochemical property distributions in drug-discovery related chemical databases is a commonly used methodology for library design [5, 15]. Fragments were characterized by widely used physicochemical descriptors such as c log P, c log D (at pH 7.4), polar surface area (PSA) (at pH 7.4), number of rotatable bonds, number of hydrogen bond acceptors, number of hydrogen bond donors, acidic dissociation constant pK_a (related to the strongest center), basic dissociation constant pK_a (related to the strongest center), number of nitrogen atoms and number of oxygen atoms using ChemAxon's JChem for Excel [16]. All the descriptors were calculated for all the fragments of both the active and the reference sets, followed by the calculation of the distribution-functions for each property. The corresponding distribution-functions were described by medians, means and standard deviations shown in Table 1.

Table 2 Classification of the acid–base characters of active and inactive fragments

	pK_a (acidic)	pK_a (basic)	Actives	Inactives
Acid	<7	>7 or absent	0	12
Base	>7 or absent	>7	83	15
Neutral	>7 or absent	<7 or absent	16	71
Zwitterion	<7	>7	1	3

The distributions of the hydrogen bond donor and acceptor-count and log P do not differ markably in their medians and standard deviations comparing the active and reference sets, however in case of the aminergic GPCR-like fragments the polar surface area (at pH 7.4) has median lower than that of the reference set. Lower log D values at pH 7.4 are related to the basic character of the aminergic GPCR fragments. Considering the number of rotatable bonds, aminergic fragments are typically less flexible containing generally 0 or 1 rotatable bond than molecules in the reference set. Fragments of the active and inactive sets were also classified by their acid–base character. The classification method and the results are shown in Table 2. 83 % of the active fragments have basic character, 16 % of the fragments were neutral, none of them were acidic, and only 1 % were zwitter-ionic. In contrast, 15 % of the randomly sampled reference fragments were basic, 71 % neutral, 12 % acidic, and 3 % zwitter-ionic. The dominantly basic character of the active set causes that although they have less O atoms, smaller PSA and slightly higher log P than the reference molecules their log D values at pH 7.4 are lower. Summarizing, these aminergic fragments constitute a set of small, rigid molecules, containing few heteroatoms, that are most often (~83 %) basic nitrogens.

Differences in property-distributions were identified by visual inspection and by the analysis of medians, means and standard deviations, comparing aminergic-fragments, and ChEMBL inactive reference fragments (see Table 1). As a result, log D (pH 7.4), number of nitrogen atoms, number of oxygen atoms, number of rotatable bonds, strongest basic pK_a, and PSA (at pH 7.4) were identified as descriptors that are able to characterize aminergic-fragments. Differences between selected property distributions were checked by Mann–Whitney U-tests [17]. All compared property distribution pairs are different at a significance level of $p = 0.05$. The following section will present some general statements about aminergic fragments, providing standpoints for developing a desirability scoring function.

Log D distribution of the amine-like fragments has a mean of 0.90, with a standard deviation of ±1.62, correlating with their predominantly basic character, differing from the log D distribution of the random fragments (mean = 1.65, SD = ±2.24), however there is no notable

Table 3 Mean, median, and standard deviation values of the calculated physicochemical properties of the active set defined by pActivity metrics

Active set ($pK_i \geq 4$)	Median	Mean	SD
PSA (at pH 7.4)	37.97	40.22	21.38
log D (at pH 7.4)	0.97	0.90	1.62
Number of nitrogen atoms	2.00	2.28	1.26
Number of oxygen atoms	1.00	1.10	0.98
Number of rotatable bonds	2.00	2.76	1.96
pK_a (strongest basic)	8.54	7.83	2.99

difference in the corresponding log P distributions. The number of nitrogen atoms represents that aminergic GPCR-like fragments have at least 1 nitrogen-atom (~40 % contain 1, ~40 % contain 2). The number of oxygen atoms puts restraint to the number of oxygens (~30 % do not contain any). Aminergic fragments are less flexible with a mean of 2.64 rotatable bonds, and with 10 % not having any rotatable bonds, unlike random fragments, with a mean of 3.35 rotatable bonds. Only 0.2 % of aminergic fragments have no basic-center. 83 % of them have at least one basic center, that is identified as having no acidic pK_a lower than 7, and a basic pK_a higher than 7. The distribution of basic pK_a has a mean value of 7.92 with a standard deviation of 2.84. In contrast, 50 % of the reference fragments are neutral (pK_a mean = 5.96, SD = 2.59). The polar surface area of a molecule is defined as the surface sum over all polar atoms. PSA is a commonly used medicinal chemistry metric for the optimization of passive membrane permeability. The widely known Rule of Three [11] filter for fragment-like compounds limits polar surface area to be at most 60 $Å^2$. The training set has a median value of 37.3 $Å^2$ and a mean value of 39.67 $Å^2$ for polar surface area, with a standard deviation of 21.04. The lower values of PSA of the active set are due to the smaller number of oxygen atoms.

The examined physicochemical properties were calculated for the fragments of the alternative active set selected by pActivity cut-off ($pK_i \geq 4$). The corresponding, medians, means and standard deviation are shown in Table 3.

Comparing property distributions based on SILE and pActivity thresholds we concluded that these are highly similar and the SILE active set was used in all further studies.

After determining the characteristic descriptors showing remarkable difference in their medians and standard deviation with respect to randomly sampled reference fragments we used these descriptors to derive a desirability scoring scheme.

Common approaches for multiparametric filtering [18] include desirability functions [19] that are used to convert preferred descriptor distributions into a scoring function. A desirability function maps the value of a property onto a score in the range of [0; 1]. Desirable and undesirable property-ranges (x) are defined by the function through series of inflection points [20], identifying the desirable and undesirable regions of the properties with a certain desirability score of y(x). Molecules not satisfying the criteria of the examined descriptor receive a desirability score of 0. Molecules with desirable properties gain a score higher than 0 and at most 1. After calculating a desirability score for all of the properties, the scores may be combined by summation or multiplication into an overall desirability measure. Desirability functions make it possible to define smooth boundaries for a property, rather than using a rigid cut-off, thus avoiding rejection of compounds based on an uncertain property value close to a criterion boundary. The selected properties were next translated to desirability functions. The two properties describing the number of heteroatoms and the number of rotatable bonds are considered to be discrete variables, while log D (at pH 7.4), PSA (at pH 7.4) and basic pK_a (strongest) are continuous (see the Supporting Information).

The number of nitrogen atoms is an important feature of aminergic GPCR-fragments due to their basic character that gives a major contribution to the interaction with the binding site. However not all fragments are supposed to be sorted out that lack nitrogen atoms. The desirable property range for the number of nitrogen atoms is between 0 and 5, namely 0 and 3–5 nitrogen atoms are scored with $y(x) = 0.5$, and 1–3 nitrogen atoms are cored with $y(x) = 1$. A score value of $y(x) = 0$ was assigned to fragments containing more than five nitrogen atoms (Figure 1 in Supporting Information).

The active fragments of the training set contained mostly 0–3 oxygen atoms, supporting the account of oxygen atoms most probably as hydrogen-bond acceptors. Fragments having 0–3 oxygen atoms are considered to be desirable with a score of $y(x) = 1$ (see in Figure 2 in Supporting Information).

In addition to the gain in conformational entropy a recent paper [11] has revealed the positive influence of the restriction of molecular flexibility by the number of rotatable bonds in ADME/PK properties suggesting an upper bound of three. The distribution of the rotatable bond count for the training set showed that ~8 % of the active fragments possessed rigid planar structure, ~21 % contained only one and ~27 % only two rotatable bonds. None of the active fragments contained more than seven rotatable bonds. Consequently all planar, non-flexible fragments and molecules containing 3 or less rotatable bonds are desirable (score of 1). Less-desirable property-ranges of 4–7 rotatable bonds are mapped with a monotonic decreasing score function (see Figure 3 in Supporting Information).

The desirability function for log D (at pH 7.4) prefers mostly fragments in the log D range of 1 and 2, and decreases monotonously towards −2 and 5, accepting more basic and more polar fragments (see Figure 4 in Supporting Information).

The desirability function of the polar surface area (at pH 7.4) is more permissive than the upper bound of the Rule of Three (PSA $\leq$ 60 Å^2), because fragments on the range of 0 and 100 Å^2 are not discarded. In contrast, these fragments are favored by a score of a hump function with a maximum at 40 Å^2 and linearly decreasing towards 0 and 100 (see Figure 5 in Supporting Information).

Though 83 % of the fragments of the training set are basic, the desirability function of the strongest basic pK_a does not totally discard an acidic fragment, in this manner the property range of the strongest basic pK_a was mapped with a constant score of 0.1 in the 0–6 pK_a range, and with a hump function covering the values of 6 and 13, penalizing fragments with basicity stronger than $pK_a > 13$, and preferring fragments with a pK_a value between 8 and 10 (see in Figure 6 in Supporting Information).

Desirability scores may be combined either by summation or by multiplication into an overall desirability measure. A disadvantage of the multiplicative approach is that it discards molecules if they receive a desirability score of zero for a single property. In contrast, the additive approach reduces the impact of a single property-score when a large number of properties are involved [18]. Taking these aspects into consideration we defined the "Fragment Aminergic GPCR Score" (FrAGS) as the sum of the individual property-scores, providing a score with a range between 0 and 6.

Validation of the Fragment Aminergic GPCR Score (FrAGS)

Validation was carried out by three complementary approaches. First, the active training set was mixed with random inactive ChEMBL compounds to check whether the score is able to sort out the active compounds. The second approach used an independent active set from the PubChem database [21]. The third validation used data of a HTS and a FS campaign on aminergic GPCR targets provided by Gedeon Richter Plc. Compounds not satisfying our fragment criterion ($8 < N_{heavy} < 22$) were removed from the HTS data set. The effectiveness of the FrAGS was determined by the enrichment factor (EF), and by the comparison of false negative rates (FNR) and true negative rates (TNR) as functions of the score. Furthermore, receiver operating characteristic (ROC) curves were also investigated.

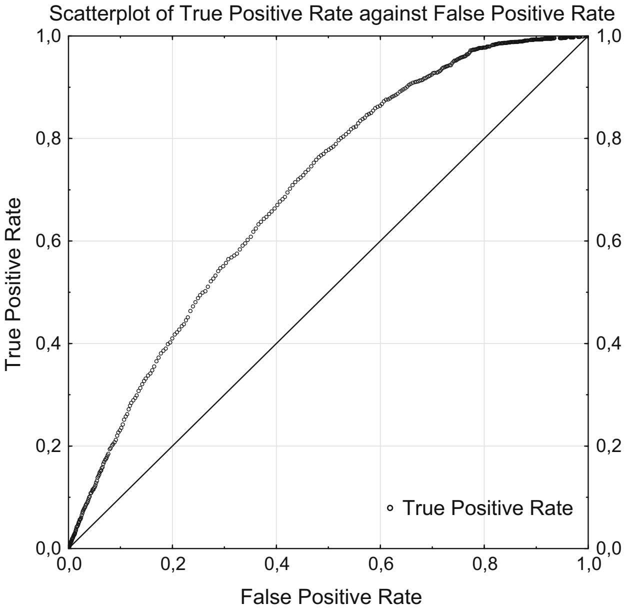

Fig. 1 Receiver operating characteristics of the ChEMBL validation data

ChEMBL validation

The inactive set was created similarly to the reference set generated for deriving the FrAGS. The entire ChEMBL database [13] of about 1,300,000 entries was filtered by the previously identified active fragments. Counter ions of salts were removed; finally 96,539 unique fragments were kept. This set was merged with the 2,183 active fragments to give a validating set with 2.2 % actives. Relevant properties were calculated for the compiled 98,722 fragments, and FrAGS were calculated. The enrichment factor increases monotonically up to the score of about 5 and oscillates for higher scores (see Figure 7 in Supporting Information). Its highest value before the oscillatory region is about 2.5. The corresponding ROC curve is shown in Fig. 1 (true negative rates and false negative rates as functions of the FrAGS are shown in Figure 8 in Supporting Information).

PubChem validation

Second validation was carried by retrieving aminergic GPCR targeted HTS data from PubChem. PubChem is a public repository for biological properties of small molecules hosted by the National Institutes of Health (NIH, USA), containing bioassay data on more than 700,000 compounds, providing information about HTS and fragment-screening data on GPCR-targets suitable for our validation [21]. High-throughput screenings for allosteric-binders, or targeting β-arrestin pathways, or used as counter-screening were sorted out. Only confirmed actives with at least 10 μM activity were taken into consideration. The

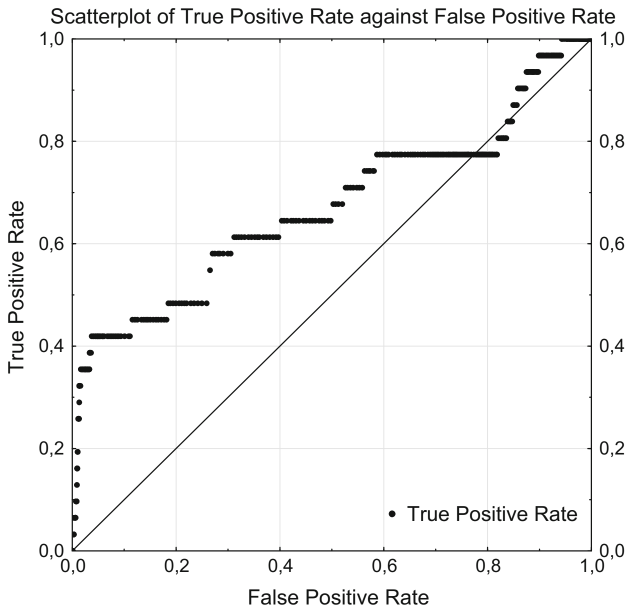

Fig. 2 Receiver operating characteristics of the PubChem validation data on $5HT_1$

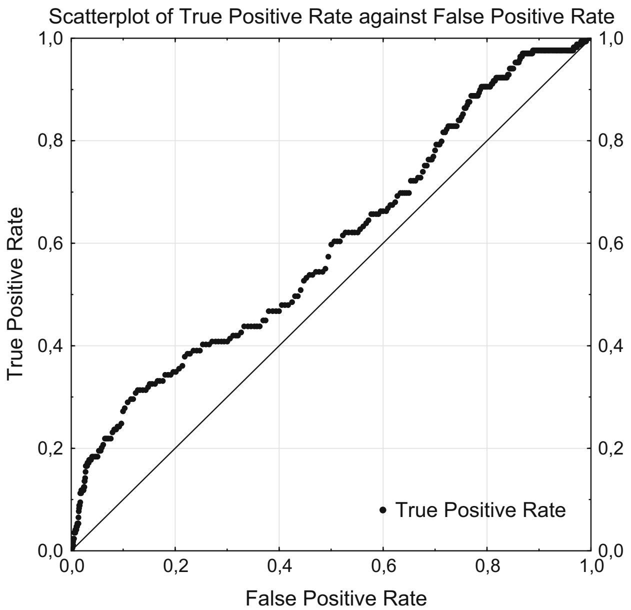

Fig. 3 Receiver operating characteristics of the PubChem validation data on $TAAR_1$

following aminergic GPCR targets were represented in HTS campaigns: $5HT_{1A}$ with 7 confirmed actives, $5HT_{1E}$ with 51 confirmed actives, $TAAR_1$ with 375 confirmed agonists and with 36 confirmed antagonists. Molecules satisfying fragment-size criterion (heavy atom count between 8 and 22) were kept providing 200 fragments with confirmed active data on aminergic-GPCRs ($5HT_{1A}$, $5HT_{1E}$, $TAAR_1$). This dataset does not overlap with ChEMBL actives used as training set. 31 confirmed $5HT_1$-active fragments and 169 $TAAR_1$-active fragments were added to a set of inactive fragments extracted from the corresponding primary screening sets by filtering all inactive compounds by the number of heavy atoms ($8 \leq N_{heavy} \leq 22$). After removing duplicates 296,510 fragments remained out of which 3,100 and 16,900 randomly sampled inactive molecules from the $5HT_1$ and $TAAR_1$ screens, respectively, were selected for validation. FrAGS values were calculated for the 3,131 $5HT_1$ set of fragments and 17,069 $TAAR_1$ set of fragments, followed by the calculation of TPR, FPR, TNR, FNR, ROC and EF (shown in Figs. 2, 3 and Table 4 and in Figures 9, 10 and 11, 12 for $5HT_1$ and $TAAR_1$ in Supporting Information). In the case of the $5HT_1$ validation the highest enrichment resulted a remarkable ratio of 7.98 ($EF_{5.0} = 7.98$). The ROC curve of the $5HT_1$ validation set (Figure 2 in Supporting Information) appears to be separated into two parts, one below and another above FPR = 0.8. It raises the question if the compounds of the two subsets belong to different structural classes. Therefore we examined the structural similarity of the two distinct sets. The distance matrices of fingerprints were calculated for the entire $5HT_1$ validation set, and separately for the two subsets. The similarity of the three sets was characterized by the means and standard deviations calculated for their pairwise distances. The mean distance was 0.66 ± 0.09, and 0.64 ± 0.09 for the first and the second subset, respectively. Furthermore the entire $5HT_1$ set has a mean distance of 0.66 ± 0.09, demonstrating the structural diversity of the validation set. The $TAAR_1$ EF curve shows a 3.41-fold enrichment at FrAGS = 5. The enrichment monotonously increases until FrAGS = 5, and oscillates over 5.

Table 4 Enrichment factors at selected FrAGS values obtained in validation studies

Fragment-GPCR-score cut-off	EF ChemBL	EF $5HT_1$	EF $TAAR_1$	EF Richter HTS	EF Richter FS
4.5	1.79	2.00	1.39	1.33	1.09
5	2.35	7.98	3.41	2.51	2.34
5.5	2.56	22.86	4.48	1.54	2.94

Validation on proprietary data

The third validation was carried out on experimental screening data of fragments, provided by Gedeon Richter Plc. The targets of the screening campaigns were aminergic GPCR receptors. Only fragment sized molecules were examined from the HTS campaign. 9,302 fragments were screened by high-throughput screening, and 3,038

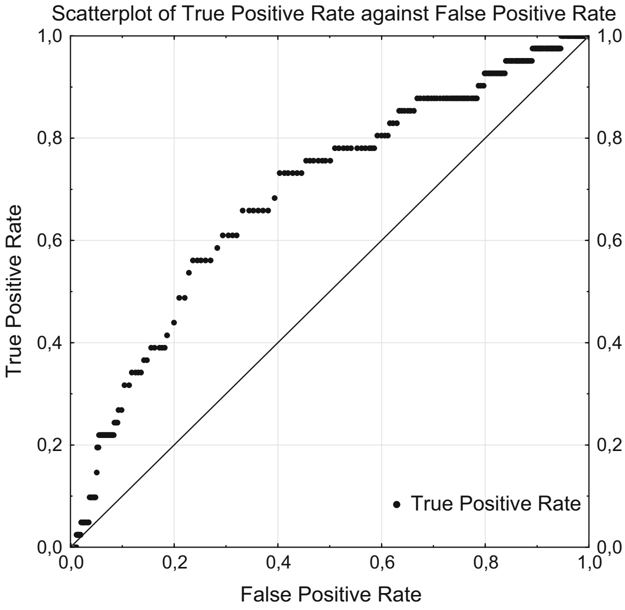

Fig. 4 Receiver operating characteristics of the GPCR HTS validation data

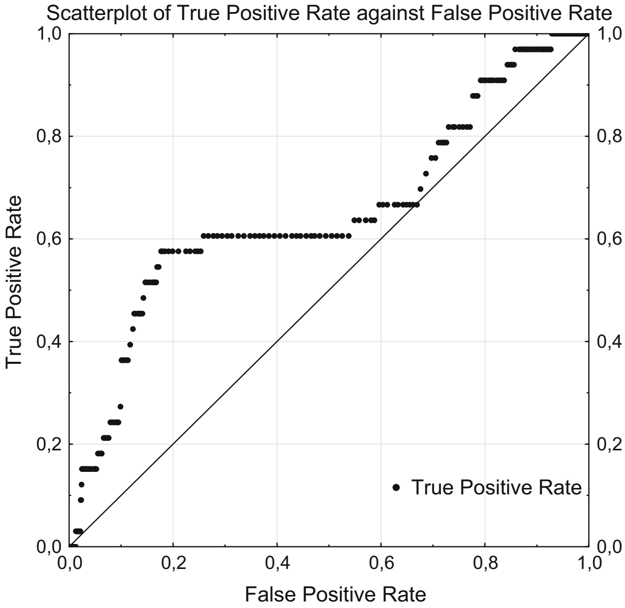

Fig. 5 Receiver operating characteristics of the GPCR FS validation data

fragments were evaluated with the fragment-screening approach.

In the high-throughput screening dataset 41 fragments were found to be active on the aminergic GPCR target, and 9,261 fragments were inactive. The enrichment factor increases monotonically until FrAGS of about 5, and oscillates over this value. The corresponding false negative rates, true negative rates and the ROC curve is shown in Fig. 4 and in Figures 13 and 14 in Supporting Information.

Fragment-screening identified 33 active, and 3,038 inactive fragments. The plot of enrichment factor versus FrAGS is shown in Figure 16 in Supporting Information. The enrichment factor increases monotonically until about FrAGS = 5. The ROC curve is shown in Fig. 5 and the plots of false negative rates and true negative rates are shown in Figures 15 and 16 in Supporting Information.

Defining FrAGS cut-off

Ligand-based filters are typically used for selecting a set of compounds of a large library for further virtual or experimental studies. For the determination of the appropriate cut-off the enrichment curves of the validation studies were examined (Table 4 and Figures 7, 9, 11, 13 and 15 in Supporting Information). The validation test using ChEMBL compounds showed, that the enrichment is about two to threefold over about FrAGS = 4.5. The enrichments are much higher in the other two validation studies using PubChem and the experimental HTS and FS data. Remarkable oscillation of the enrichment was observed over FrAGS > 5 that is due to the small number of compounds having high score values. It is proposed that a cut off value of 5.0 is an appropriate choice for the selection of promising aminergic fragments. Our three independent validation studies show enrichments around or above 3 using this cut off value.

Conclusions

Comparing the distributions of several physicochemical descriptors calculated for fragment-like aminergic GPCR ligands with those of randomly selected fragments revealed that the following descriptors are markedly different for the two sets of compounds: log D (at pH 7.4), PSA (at pH 7.4), pK_a (strongest basic center), number of nitrogen atoms, number of oxygen atoms, and the number of rotatable bonds. Desirability functions based on the statistical differences in the property distributions were defined and were combined to a Fragment Aminergic GPCR Score (FrAGS). This score can take values between 0 and 6, the higher the score the more likely the fragment is an aminergic GPCR ligand. The scoring scheme was verified with independent experimental data collected in real-life screening situations and it was found that a score cut off value of 5 is appropriate to achieve an enrichment of around or over 3. The scoring scheme is suitable for screening large libraries of fragment-like compounds for aminergic GPCR ligands, and thus it is a useful tool for compiling focused fragment libraries for drug discovery projects.

Supporting Information

The desirability functions, Enrichment Factors and the TNR/FNR plots are available free of charge via the Internet at http://www.springer.com/.

Acknowledgments We thank for Gedeon Richter Plc for providing GPCR-related fragment-screening and high-throughput screening data for the validation studies and particularly Márton Vass for helpful discussions.

References

1. Fink T, Bruggesser H, Reymond JL (2005) Virtual exploration of the small-molecule chemical universe below 160 daltons. Angew Chem Int Ed 44:1504–1508
2. Bohacek RS, McMartin C, Guida WC (1996) The art and practice of structure-based drug design: a molecular modeling perspective. Med Res Rev 16:3–50
3. Leeson PD, Springthorpe B (2007) The influence of drug-like concepts on decision-making in medicinal chemistry. Nat Rev Drug Discov 6:881–890
4. Hann MM, Keserű GM (2012) Finding the sweet spot: the role of nature and nurture in medicinal chemistry. Nat Rev Drug Discov 11:355–365
5. Andrews SP, Brown GA, Christopher JA (2014) Structure-based and fragment-based GPCR drug discovery. ChemMedChem 9:256–275
6. ChEMBL GPCR SARfari home page. https://www.ebi.ac.uk/chembl/sarfari/gpcrsarfari
7. Knime Desktop (Konstanz Information Miner), version 2.9.1 (2014)
8. Willem J, Nissink M (2009) Simple size-independent measure of ligand efficiency. J Chem Inf Model 49(6):1617–1622
9. Hann MM, Leach AR, Green DVS, Oprea TI (eds) (2005) Methods and principles in medicinal chemistry, vol 23. Wiley-VCH, Weinheim, pp 43–57
10. Hopkins AL, Groom CR, Alex A (2004) Ligand efficiency: a useful metric for lead selection. Drug Discov Today 9:430–431
11. Congreve et al (2003) A 'rule of three' for fragment-based lead discovery. Drug Discov Today 8:876–877
12. GPCRDB information system for G protein-coupled receptors home page. http://www.gpcr.org/7tm/
13. ChEMBL home page. ftp://ebi.ac.uk/pub/databases/chembl/ChEMBLdb/
14. RDKit: cheminformatics and machine learning software home page. http://www.rdkit.org/
15. Balakin KV, Tkachenko SE, Lang SA, Okun I, Ivashchenko AA, Savchuk NP (2002) Property-based design of GPCR-targeted library. J Chem Inf Comput Sci 42:1332–1342
16. Marvin, version 5.2, JChem for Excel (2014) Chemaxon, Budapest
17. Statistica 12, Statsoft home page. http://www.statsoft.com/
18. Segall MD (2012) Multi-parameter optimization: identifying high quality compounds with a balance of properties. Curr Pharm Des 18:1292–1310
19. Harrington EC (1965) The desirability function. Ind Qual Control 21:494–498
20. Wager TT, Hou X, Verhoest PR, Villalobos A (2010) Moving beyond rules: the development of a central nervous system multiparameter optimization (CNS MPO) approach to enable alignment of drug-like properties. ACS Chem Neurosci 1:435–449
21. PubChem home page. https://pubchem.ncbi.nlm.nih.gov/
22. Tanimoto distance as defined in the "Distance Matrix Calculate" node of Knime