

GPCR-Targeted Library

Medicinal and Computational Chemistry Dept., ChemDiv, Inc., 6605 Nancy Ridge Drive, San Diego, CA 92121 USA, Service: +1 877 ChemDiv, Tel: +1 858-794-4860, Fax: +1 858-794-4931, Email: ChemDiv@chemdiv.com

Introduction

Owing to historic inefficiency of mass random bioscreening, the current paradigm suggests that target-specific and pharmacokinetic properties of small molecule libraries should be addressed as early as possible in the discovery process. Computational medicinal chemistry can address this problem at the level of pre-synthetic library design. A number of advanced *in silico* methods have recently been developed and applied to combinatorial templates to enhance their target-specific informational content. Appropriate strategies for the design of combinatorial libraries are developed in accordance with the target, disease area, resources on hand and the specific project goals.

In this description, we present a rational, practical approach to the design of GPCR-targeted combinatorial library. The goal of the combinatorial synthesis planning strategy presented here is to construct an algorithm utilizing simple, automated procedures for designing combinatorial libraries that are expected to show GPCR-activity.

1. GPCRs as promising drug target

The superfamily of G-protein-coupled receptors (GPCRs) is a diverse group of transmembrane proteins crucial to eukaryotic cellular signalingⁱ. GPCRs initiate cascades of cellular responses to diverse extracellular mediators and are involved in all common human diseases. Nearly 40% of marketed drugs act through modulation of GPCR functionsⁱⁱ and up to 70% of novel therapeutics in development target known GPCRsⁱⁱⁱ. In addition, several hundred “orphan” GPCRs (which have no natural ligands identified as yet), are the focus of an intense drug discovery effort in many programs. Characterization of orphan GPCRs will substantially facilitate research in human physiology and pharmacology. GPCR family consists of seven basic classes: Rhodopsin- and Secretin-like receptors, Metabotropic glutamate and Fungal pheromone receptors, cAMP receptors, Ocular albinism proteins and Frizzled/Smoothed subfamily. All the classes listed are additionally subdivided into several categories, for example, Rhodopsin-like GPCRs include amine subclass (Muscarinic acetylcholine, Dopamine, Histamine, Serotonin, Bradykinin, Adrenoceptors, etc.), peptide subclass (Angiotensin, Chemokine, Endothelin, Neurotensin, Opioid, Somatostatin, Tachykinin, Vasopressin-like, etc.), Nucleotide-like receptors (Adenosine and Purinoceptors) and other subclasses^{iv}.

The key to harnessing the clinical potential of particular GPCRs lies in the ability to elucidate their tissue- and disease-specific functions and identify the selective ligands for these receptors. The optimal ligands would be the potent small molecules with ADME/Tox properties required for the orally available drugs. Specifically, the optimal ligands need to possess high affinity and specificity for the target protein, and reasonable membrane permeability for biological activity in whole cell assays and *in vivo* models. The prime source of drug candidates is the focused small molecule libraries developed against particular receptors which are often compiled into protein class “GPCR-targeted libraries”. Such libraries are being built by drug discovery companies in house and are available commercially from medicinal chemistry companies. Due to significant diversity of natural GPCRs ligands and the complexity of downstream events of GPCR signaling, the optimal choice of GPCR library construction strategy represents a non-trivial and highly important problem.

2. Neural networks in the design of GPCR-targeted library

There are several approaches to the design of GPCR-focused compound libraries, ranging from 2D simulation algorithms to the analysis of ligand receptor spatial arrangement and neural network (NN) learning QSAR systems. Over the last years, the methods based on neural networks became popular due to their efficiency in solving the problem. Several recent studies described successful employment of neural network methods for segregation of pharmaceutical compounds in categories based on different properties^v. Recently, we have applied NN classification methodology for property-based design of GPCR-targeted library^{vi}. In particular, we have found that a proper combination of specific physicochemical allows to successfully differentiating GPCR ligands from compounds active against other target-specific classes. Using these findings, the NN classification models were created with excellent discriminatory power. We have also attempted to solve the next level, more difficult problem: differentiation between specific classes of GPCR ligands. The key goal of ^{vii} was to develop *in silico* procedure for the design of small-molecule libraries that would show a receptor-specific GPCR activity. In the fundamental work ^{viii} we have comprehensively investigated and reviewed peptidergic G-protein coupled receptors (pGPCRs), their small-molecule modulators as well as the related structure-based design of such agents.

Fundamentally, neural network (NN) modeling allows optimizing a large number of input parameters in different areas of NN applications. In drug development this property of NNs is used in "property-based design" approach, by analogy with the terminology proposed earlier^{ix}. NN approach is an efficient tool for constraining the size of virtual compound libraries designed for primary bioscreening with target-specific activity. The property-based approach is an alternative to a variety of more broadly used target- and ligand-structure focused design methods. Despite of its track record of success for certain targets, target-focused design has serious drawbacks. Namely, these are an inability

to accurately estimate all target-ligand interactions, significant computation time, the ignorance of water microenvironment, the difficulties in correct generation of 3D structures and in the analysis of all possible spatial conformations of it, etc. Ligand structure-based methods are indispensable in exploring the feasible chemistry space when many ligands for a

target are known and the active chemotypes are defined. However, the method is poorly suitable for the discovery of novel lead chemotypes. It is well documented that most popular ligand structure-based methodologies (such as bioisosteric approach, or similarity-based methods) are skewed toward the old scaffolds. In general case, the target- and ligand structure-based technologies can not adequately address all the real problems of rational drug design, particularly those connected with the virtual screening of large compound databases or with the discovery of novel lead chemotypes. A similarity of molecular physico-chemical properties represents an alternative design basis for target specific libraries. The underlying theory

states that every group of active ligand molecules can be characterized by a unique combination of physico-chemical parameters differentiating it from other target-specific groups of ligands. As a rule, receptors of one type share the structurally conserved ligand binding site. The structure of this site dictates the bundle of properties a receptorselective ligand should possess, such as specific steric, lipophilic, H-binding, and other features influencing the pharmacodynamic requirements. This theory is realized in computation models for quantitative discrimination between the ligand groups. Whenever a large set of active ligands is available for a particular receptor, the mean values of some key molecular properties can be considered as optimal and characteristic of this group of ligands. Based on these values, one can generate a quantitative discrimination function that permits the selection of a series of compounds to be assayed against the target. Finding such function is a key element for computational virtual screening programs. It is important for this function to be based on physico-chemical rather than on structural properties to be capable of suggesting novel lead chemotypes.

2.1. Unsupervised Kohonen-based learning approach

In most studies on application of neural networks in drug discovery, a supervised learning strategy was used. The alternative unsupervised learning method becomes popular for comparative analysis and visualization of large ligands data sets^x. For instance, benzodiazepine and dopamine data sets were compared recently with an implementation of a Kohonen network^{xi}. In another study, a dataset of 31 steroids binding to the corticosteroid binding globulin (CBG) receptor was modeled^{xii}. Kohonen self-organizing maps were used for distinguishing between drugs and non-drugs with a set of descriptors derived from semi-empirical molecular orbital calculations^{xiii}. It was emphasized that Kohonen map-based classification does not depend on the definition of a non-drug, non-ligand data set, and, therefore, the virtual screening of active compounds can be conducted more objectively. This property of

unsupervised Kohonen learning strategy is particularly important in cases when the negative training set is unavailable or hard to define. In this work, we used the unsupervised learning methodology for differentiation between various receptor-specific groups of GPCR ligands. With the data available, only positive training selections of molecules can be unambiguously identified, namely, the groups of ligands to particular GPCRs. The definition of a negative training set would be very complicated and, probably, unreliable, as only a few compounds with particular receptor-specific activity have been tested against all groups of GPCRs. This limitation restricts the application of multi-layer neural networks with a supervised learning procedure as an error back-propagation learning algorithm; an unsupervised approach is required. Among the unsupervised methods, we chose Kohonen neural network as the one with the most appropriate learning strategy for GPCR-targeted library design.

3. Concept and Applications

GPCR-targeted library design at CDL involves:

- *A combined profiling methodology that provides a consensus score and decision based on various advanced computational tools:*

1. Unique bioisosteric morphing and funneling procedures in designing novel potential GPCR ligands with high IP value. We apply CDL's proprietary ChemosoftTM software and commercially available solutions from Accelrys, MOE, Daylight and other platforms.
2. Neural Network tools for target-library profiling, in particular Self-organizing Kohonen maps, performed in SmartMining Software. We have also used the Sammon mapping and Support vector machine (SVM) methodology as more accurate computational tools to create our GPCR-focused library.
3. In several cases we have used 3D-molecular docking approach to the focused library design.
4. Computational-based *in silico* ADME/Tox assessment for novel compounds includes prediction of human CYP P450-mediated metabolism and toxicity as well as many pharmacokinetic parameters, such as Brain-Blood Barrier (BBB) permeability, Human Intestinal Absorption (HIA), Plasma Protein binding (PPB), Plasma half-life time ($T_{1/2}$), Volume of distribution in human plasma (V_d), etc.

A general approach to limiting the space of virtual libraries of combinatorial reaction products consists of implementation of a series of special filtering procedures. The typical filtering stages are briefly summarized in Figure 1. A variety of "Rapid Elimination of Swill" (REOS) filters is used to eliminate compounds that do not meet certain criteria^{xiv}.

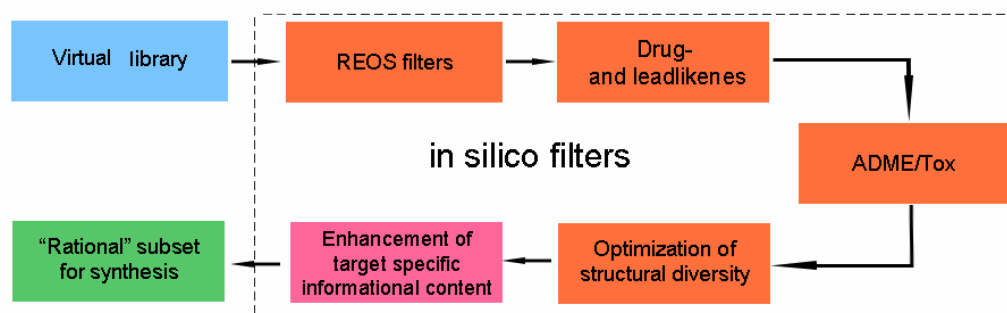


Figure 1. General procedures of selection of a rational target-specific subset within an initial virtual combinatorial library

These criteria can include: (1) presence of certain non-desirable functional groups, such as reactive moieties and known toxicophores; (2) molecular size, lipophilicity, the number of H-bond donors/acceptors, the number of rotatable bonds. At the next stage the design focuses on “lead” and “drug-likeness” of combinatorial molecules^{xv}. The ADME/Tox properties of screening candidates should be taken into consideration as early as possible^{xvi}. Additional filters are therefore used for *in silico* prediction of some crucial ADME/Tox parameters, such as solubility in water, logD at different pH values, cytochrome P450-mediated metabolism and toxicity, and fractional absorption. Optimization of structural diversity is another natural and very important way to constrain the size of combinatorial libraries (reviewed in ^{xvii}). The fundamentals for these applications are described in a series of our recent articles on the design of exploratory small molecule chemistry for bioscreening [for related data visit ChemDiv, Inc. online source: www.chemdiv.com]. Our multi-step *in silico* approach to GPCR-focused library design is schematically illustrated in Fig. 2.

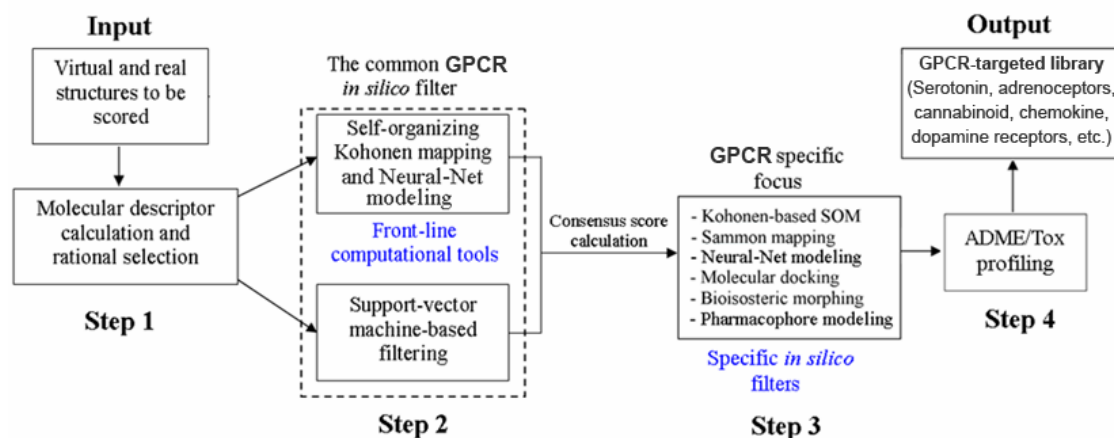


Figure 2. Multi-step computational approach to GPCR-targeted libraries design

This common approach was effectively applied for the developing of our GPCR-focused, in particular for Serotonin, Dopamine, Opioid, Endothelin, Cannabinoid, Bradykinin, Chemokine receptors, Adrenoceptors, etc.

• *Synthesis, biological evaluation and SAR study for the selected structures:*

1. High-throughput synthesis with multiple parallel library validation. Synthetic protocols, building blocks and chemical strategies are available.
2. Library activity validation via bioscreening; SAR is implemented in the next library generation.

3.1. GPCR-related reference database

Our reference set included 12639 GPCR-active agents with experimentally shown activity against over 100 different GPCRs (Table I). The set contained compounds from different stages of clinical trials, marketed drugs and patented NECs, sourced from Ensemble database of known pharmaceutical agents compiled from the patent and scientific literature^{xviii}. Molecules were filtered based on molecular weight range (150-700) and atom type content (only C, N, O, H, S, P, F, Cl, Br, and I allowed).

Table I. Reference database of GPCR ligands

| <i>Number</i> | <i>GPCR targets</i> | <i>Ligand type</i> | <i>Compounds</i> |
|---|---|----------------------|------------------|
| 1 | 5-HT1A/1B/1D | Agonists/antagonists | 1749 |
| 2 | 5-HT2A/2B | Agonists/antagonists | 463 |
| 3 | 5-HT4 | Agonists/antagonists | 164 |
| 4 | α 1/ α 2-Adrenoceptor | Agonists/antagonists | 713 |
| 5 | β -Adrenoceptor | Agonists/antagonists | 1000 |
| 6 | Bradykinin B2 | Agonists/antagonists | 126 |
| 7 | Cannabinoid CB1/CB2 | Agonists/antagonists | 93 |
| 8 | CCKB | Antagonists | 200 |
| 9 | CRF | Antagonists | 372 |
| 10 | δ -Opioid | Agonists | 195 |
| 11 | Dopamine Autoreceptor | Modulators | 189 |
| 12 | Dopamine D1-D4 | Agonists/antagonists | 1551 |
| 13 | Endothelin ETA/ETB | Antagonists | 319 |
| 14 | κ -Opioid | Agonists | 164 |
| 15 | μ -Opioid | Agonists/antagonists | 86 |
| 16 | Muscarinic M1 | Agonists | 630 |
| 17 | Neuropeptide Y | Antagonists | 101 |
| 18 | Oxytocin | Antagonists | 209 |
| 19 | PGE2 | Antagonists | 129 |
| 20 | Tachykinin NK1/NK2 | Antagonists | 1734 |
| 21 | Vasopressin V1/V2 | Antagonists | 126 |
| 22 | Approximately 80 minor GPCR-specific groups with number of compounds < 50 | Agonists/antagonists | 2700 |
| Total number of compounds in training set ^{a)} | | | 12 540 |

N.B. 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

Diversity parameters for this reference database are shown in Table II. As evident from the number of screens, the number of core heterocyclic fragments, and the diversity coefficients (all these parameters are calculated using the *Diversity* module^{xix} of the ChemoSoftTM software tool), the studied compound database has high structural diversity and can be considered to be a good representation of known GPCR-active compounds.

Table II. Diversity parameters of the studied database

| Parameter | Value |
|------------------------------------|-------|
| Total number of compounds | 12540 |
| No. of screens ^a | 13503 |
| Diversity coefficient ^b | 0.822 |
| No. of core heterocycles | 1131 |

^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 non-diagonal similarity matrix elements are used in ChemoSoftTM program as a diversity measure; the diversity coefficient can possess the value from 0 to 1, which correspond to minimal and maximal possible diversity of a selection.

3.2. Molecular descriptors

The Kohonen-based model was based on a pre-selected set of molecular descriptors. Sixty molecular descriptors describing the important molecular properties, such as lipophilicity, charge distribution, topological features, steric and surface parameters were explored. The number of descriptors was reduced by the omission of the low-variable and highly correlated ($R > 0.9$) descriptors. To further reduce the descriptor space, we performed a principal component analysis using SmartMining software suite^{xx}. Eventually, seven descriptors were selected as the most relevant and further used in the neural network experiments (Table III). The chosen descriptors are readily computable and, in combination, provide a reasonable basis for the assessment of the particular GPCR activity potential. This set of descriptors defines a bundle of most relevant factors affecting the ability of a compound to possess GPCR-activity: lipophilicity, molecular surface area and size, H-binding potential and surface charge properties.

Table III. Molecular descriptors used for modeling

| Descriptors | Definition |
|-------------|---|
| MW | molecular weight |
| logD | log of partition coefficient in 1-octanol/water at pH 7.4 |
| HBD | number of H-bond donors |
| HBA | number of H-bond acceptors |
| Rot-B | number of rotatable bonds |
| TPSA | total polar surface area |
| PPSA-1 | partial positive surface area |

3.3. Kohonen map generation

In this work we have used the internally developed program included in CDL's proprietary ChemoSoft™ software suite, for unsupervised learning and generation of Kohonen maps. A 15x15 node architecture was chosen in order to provide the studied molecules with optimal distribution space. The reference database was used for NN-training and Kohonen map generation (Figure 3).

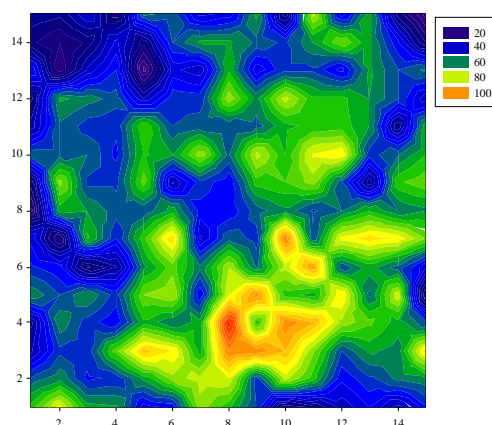


Figure 3. 15x15 Kohonen network trained with seven selected descriptors for total reference database of GPCR-active agents (12K cmpds). The data have been smoothed

As shown in figure 3, the GPCR ligands are widely distributed throughout the map as the irregularly shaped islands, with a trend towards the bottom of the map. The area occupied by the GPCR ligands is relatively large, which reflects their significant diversity. At the next step, we studied the distribution of different receptor-specific groups of GPCR ligands within the generated Kohonen map. These ligand groups appeared to be clustered at different distinct areas of the map. As an illustration, Figure 4a-d shows the distributions of four large GPCR-specific ligand groups. Interestingly, the active agents entering into clinical trials or launched drugs, usually gravitate towards the central parts of the corresponding receptor-specific sites on the map, while the peripheral positions are occupied by

compounds at more earlier stages of development (data not shown). The Kohonen maps for particular receptor-specific groups of ligands can be used for predicting potential receptor-specific activity. Thus, the processing of a diverse exploratory compound library on this Kohonen map allows to distinguish between the specific compound subsets falling into particular receptor-specific areas. We suggest that the molecules from these subsets are more likely to be active against the corresponding receptors.

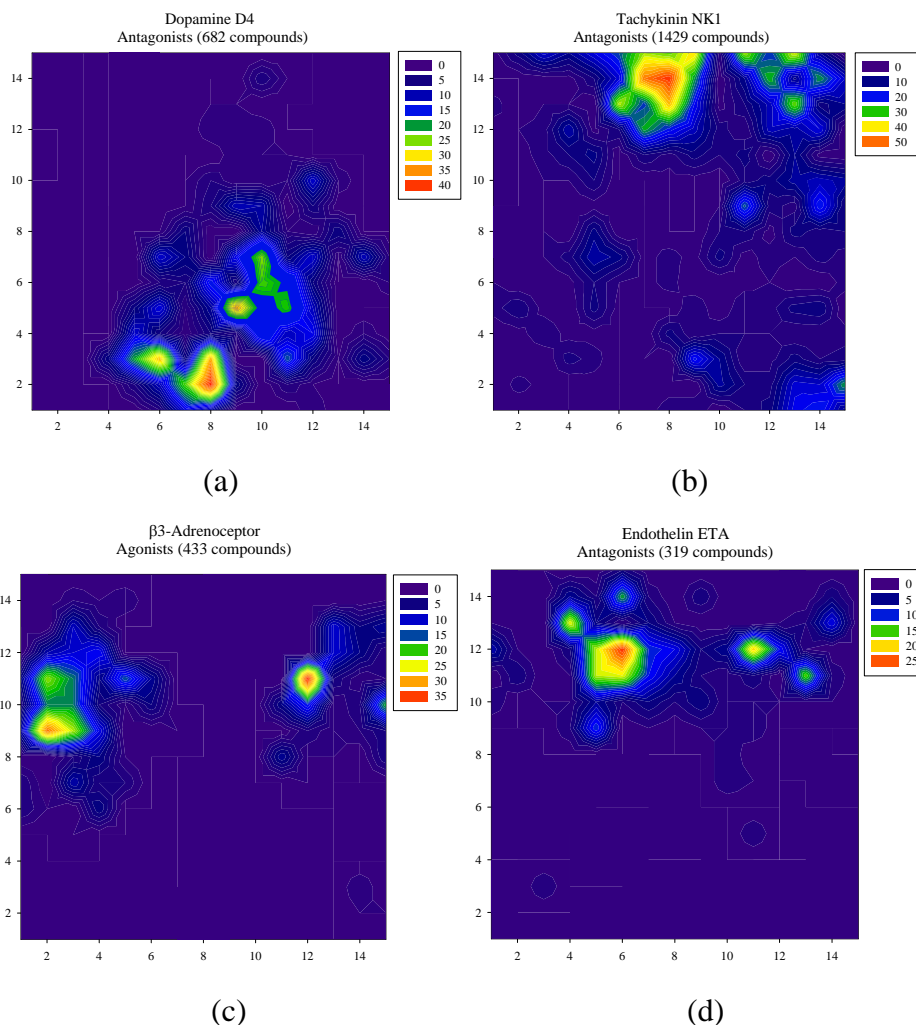


Figure 4. Distributions of four large GPCR-specific ligand groups within the Kohonen map. Other GPCR-groups, including Serotonin, Muscarinic M1, Opioid and Chemokine receptor antagonists, were also found to locate in a separate field within the map constructed, but they are not shown here

Three-dimensional distribution plots of these ligand groups (Figure 5a-c) demonstrate statistically significant differences in their location.

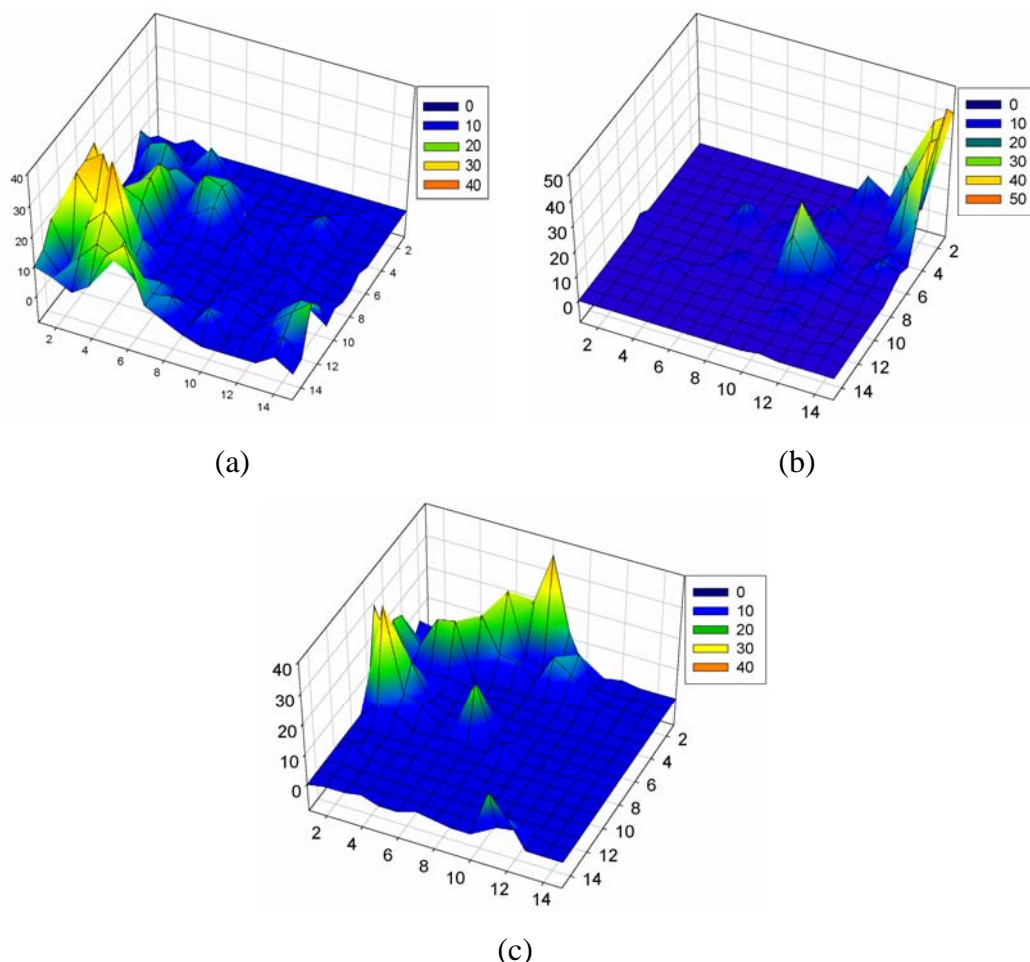


Figure 5. 3D diagrams of distribution of three target-specific groups of GPCR ligands on the Kohonen map: (a) Tachykinin NK1 Antagonists, (b) Muscarinic M1 Agonists, β 3-Adrenoceptor Agonists

3.4. Rational design of GPCR-specific combinatorial libraries based on the concept of privileged substructures

As mentioned above, "GPCR-activity" assumed here as the ability of a small-molecule compound to be a successful ligand for a GPCR. Thus, within this section we clearly demonstrate the practical significance of the concept of privileged substructures in the identification of combinatorial building blocks for synthesis of GPCR-focused library rich in target-specific structural motifs. We also illustrate the novel and interesting possibilities associated with the property-based selection of reactants and products, using advanced machine learning strategies that have been developed to identify structural motifs which specifically interact with biotargets or target families using retrosynthetic analysis of existing knowledge bases and generation of specific molecular fragments^{xxi}. Thus, a RECAP (Retrosynthetic Combinatorial Analysis Procedure) method was described based on fragmenting

molecules around bonds which are formed by common chemical reactions^{xxii}. The main advantage of this approach is that the initial molecules are fragmented at several predefined bond types, all of which are amenable to combinatorial chemistry. Therefore, the resulting fragments represent direct precursors of building blocks for combinatorial library synthesis. Though the RECAP technique is very useful in the design of combinatorial libraries, this pure retrosynthetic approach can lead to ungrounded simplification of some privileged scaffolds. For example, according to RECAP rules, biphenyl fragment is dissected. We have used a modified approach which takes into consideration not only the chemistry-derived rules but the distinctive structural features of some GPCR privileged scaffolds. Along with the cleavage rules, we specified several bond types, which are left intact. Thus all mono- and biheterocyclic structures, benzylheterocycles, spirocyclic fragments, biphenyl and diarylmethane fragments and their heterocyclic bioisosteres, as well as all ring fragments are considered as indestructible.

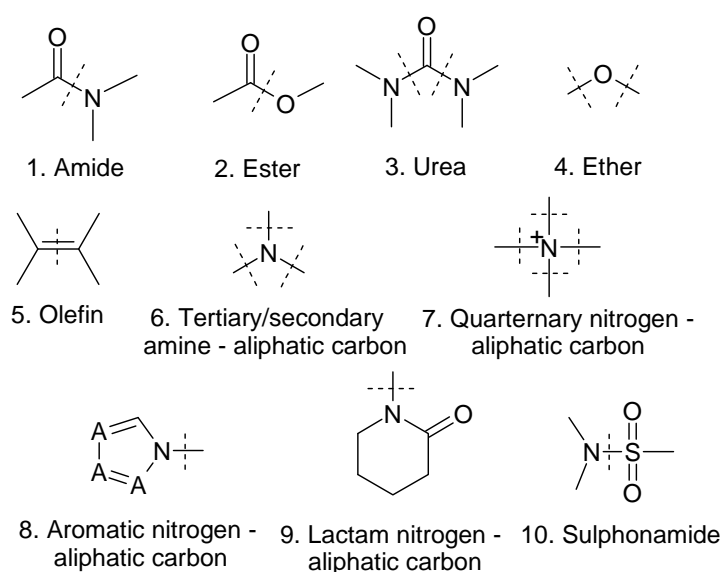


Figure 6. Main chemical bond cleavage types

The main chemical bond types at which to cleave a molecule are shown in Figure 6. Hydrazides, hydrazones, ketoximes, ureas, uretanes, esters of hydroxamic acids are cleaved in the similar manner. If the terminal fragment to be cleaved contains only small functional groups with molecular weight less than 45, the fragment is left uncleaved. The non-terminal cleaved fragments with molecular weight less than 45 are eliminated. The main reasons for this are to avoid generating very simple fragments, and to obtain more "drug-like" fragments. An example of a typical cleavage is given in Figure 7, where an initial molecule, Alfuzosin, with three cleavage points and three resulting fragments are shown.

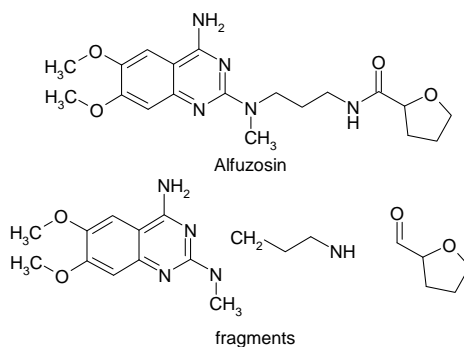


Figure 7. Results of dissection of α 1-adrenoceptor antagonist Alfuzosin

It is important to note that the applied rules are not the pure retrosynthetic rules. The described structure dissection procedure also takes into account the structures of typical privileged motifs and is directed to the search of structural chemotypes with selective action against particular receptors. As mentioned in the previous section, our method is based on extracting information from large reference databases of active agents which show any target-specific activity. The reference database of GPCR-active agents (see section 3.1) was then fragmented using a structure splitting package of the ChemoSoftTM software and the cleavage rules described above. The procedure is automatic for each particular rule, and it results in a database of final fragments for which no further cleavage is possible. In this database, target-specific activity of a parental molecule is indicated for each fragment. Table IV shows the numbers of fragments obtained from the initial database and for three arbitrary receptor-specific groups of ligands.

Table IV. The number of fragments obtained as a result of structure dissection procedure for the initial reference database

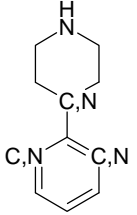
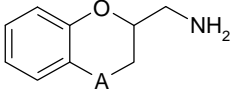
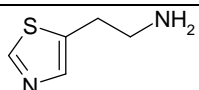
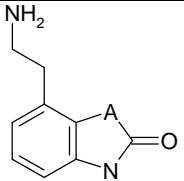
| | Fragments | Unique fragments |
|--|------------------|-------------------------|
| total fragmented database | 32756 | 7425 |
| fragmented set of Dopamine D2 agonists | 1839 | 452 |
| fragmented set of Tachykinin NK1 antagonists | 5211 | 1400 |
| fragmented set of Cannabinoid agonists/antagonists | 294 | 120 |

We also have shown that the privileged substructures, which are likely associated with target-specific activity of the uncleaved compounds, are present in fragmented GPCR-specific groups of ligands. As an example, we have used the fragmented database of Dopamine D2 agonists. This database, obtained by fragmenting 89 intact Dopamine D2 agonists, contains 1839 non-unique and 452 unique final retrosynthetic fragments. It was clustered using a method based on the distance matrix derived from Tanimoto similarity coefficients. As a result, several clusters containing more than five structures were generated, which represent the most frequently occurring fragments in the Dopamine D2 database.

To correctly address the problem of identification of target-specific privileged motifs, one should take into account the phenomenon of bioisosterism^{xxiii}. Thus several different bioisosteric structures can constitute only one distinct privileged structural motif. In order to include all possible bioisosteric analogs into one cluster, we have used a special algorithm of ChemoSoftTM based on a collection of rules for bioisosteric conversions described in literature. All bioisosteric analogs are considered similar with similarity coefficient 1 if they have identical substituents around the central bioisosterically transformed fragment. To facilitate analysis of the association of specific fragments (possible privileged motifs') with a given target-specific dataset, we constructed a *characteristic occurrence* metric. For each privileged motif obtained after the cleavage procedure, we determined its occurrence in each GPCR-specific data set, and then compared this to the frequency of occurrence in the entire fragmented database. To construct the characteristic occurrence (CO) metric for a fragment in a particular set, we calculated the percentage ratio of the fragment's occurrence in the set to the total number of compounds in this set. To quantitatively assess an enrichment of a particular activity set with a specific fragment, we used the CO of this fragment in a particular compound set relative to its CO in the whole fragmented data set. The ratio of these characteristic occurrences of any fragment can serve as a measure of uniqueness of the fragment's distribution in the corresponding receptor-specific fragment base compared to the total database.

As an illustration, Table V shows four heterocyclic structures that were present in the two data sets, the fragmented database of Dopamine D2 receptor antagonists and the total fragmented database, with at least five-fold difference in the CO values, $CO_{Dop_D2}/CO_{tot} > 5$. Such a difference gives a reasonable indication of whether a fragment is specific for this particular activity group or whether it is widely distributed in many unrelated groups. In fact, the fragments shown in Table V represent privileged substructural motifs for Dopamine D2 receptor antagonists.

Table V. Some privileged structural motifs of Dopamine D2 agonists

| Fragment | CO _{tot} , % | CO _{Dop_D2} , % | CO _{Dop_D2} / CO _{tot} |
|---|--------------------------|-----------------------------|---|
|  | 0.42 | 10.18 | 24.2 |
|  | 0.59 | 3.34 | 5.7 |
|  | 0.21 | 2.41 | 11.5 |
|  | 0.13 | 6.21 | 47.8 |

In a similar manner, such privileged motifs can be identified for each GPCR-specific activity group. The typical number of privileged substructures per group lies in the range of 5-20 for the studied GPCR-specific compound sets.

The *N*-arylpiperazine fragment shown in Table V represents an interesting structural motif with an expressed mixed type of receptor-specific activity. Compounds containing this fragment can be active against Dopamine D2, Tachykinin NK1, Vasopressin V1A/V2 and other GPCRs (see also Table VI in the following section). Such activity with respect to the entire GPCR family substructures represents very valuable objects in combinatorial synthetic strategies. Their ability to serve as selective ligands against different receptors can be modulated with the proper selection of other parts of the molecule.

3.5. Privileged versus peripheral retrosynthetic fragments

Analysis of GPCR ligands allows us to identify two principal categories of retrosynthetic fragments. The main category is the privileged fragments, described in the previous section. In most cases, the occurrence of a privileged motif is crucial for the target-specific activity of a compound. It should be noted that identification of privileged target-specific motifs is rather a technical problem, in the sense that whenever a large enough reference database of active agents and the appropriate chemical

database management tools, are available (such as chemical database, clustering package module, bioisosteric similarity module etc.) privileged target-specific substructures can be identified.

Peripheral structural motifs are a second principal category of retrosynthetic fragments. The presence of such fragments usually does not substantially influence the target specificity of a compound (with exception of molecules containing the privileged core motifs with multiple target selectivity), but can seriously affect the protein-ligand binding affinity, as well as its pharmacokinetic properties. In contrast with privileged motifs, the peripheral fragments belonging to a particular target activity class usually do not possess any structural similarity. Thus several selective Dopamine D2 agonists have a distinct privileged substructure motif (*N*-arylpiperazine moiety) and different peripheral fragments highlighted in blue (Figure 8).

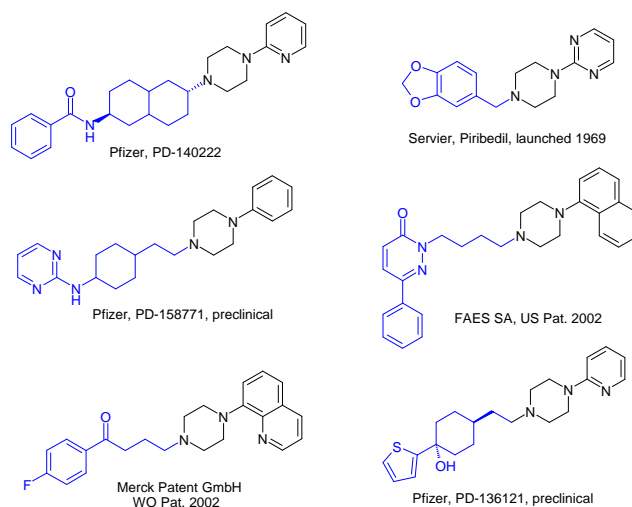


Figure 8. Dopamine D2 agonists having a distinct *N*-arylpiperazine privileged motif and structurally different peripheral fragments

Nevertheless, the structural peripheral fragments shown in Figure 8 have similar size, molecular topology, lipophilicity, number of H-bond accepting groups and number of rotatable bonds. It can be concluded that peripheral structural motifs, being structurally different molecular fragments, can exhibit similar physico-chemical and spatial properties.

3.6. Peripheral retrosynthetic fragments: How to measure the target-specific differences?

In stage of our study, we seek to answer the following important question: whether we can quantitatively discriminate between different target-specific combinations of peripheral cleaved fragments based on their molecular properties, rather than on their structural dissimilarity. To address this question, we also used an advanced data mining method based on artificial neural networks and unsupervised learning approach. Thus, we have successfully applied an advanced method of NN quantitative SAR and data visualization based on Kohonen self-organizing maps. A similar strategy was recently used and described in our work devoted to prediction of cytochrome P450-mediated metabolism of organic compounds^{xxiv}. In these experiments, we used the entire 7452-compound

database of unique cleaved fragments. Each fragment in this database is characterized by a defined profile of target-specific activity of its active compound-precursor, focused against 1 of more than 100 different GPCR targets. Molecular features encoding the relevant physicochemical properties of compounds were calculated from 2D molecular representations of the molecular fragments. Fragment size, topological complexity, H-binding capacity and lipophilicity were the main contributors to the models generated.

A self-organizing Kohonen map of the total database of cleaved retrosynthetic fragments generated as the result of an unsupervised learning procedure (data not shown), indicates that the cleaved fragments occupy a wide area on the map, characterized as the area of potential building blocks for combinatorial synthesis. Studying the distribution of various target-specific groups of peripheral structural fragments in the Kohonen map yielded interesting results consistent with our intuitive hypothesis about similarity of physico-chemical properties between peripheral retrosynthetic fragments belonging to a particular target-specific category. Most of the groups have distinct locations in specific regions of the map (Figure 9a-e). The differences in location sites allow us to formulate the underlying principle, which can be used for selection preferred peripheral fragments for each particular receptor-specific category: every group of peripheral structural fragments associated with defined target specificity can be characterized by a distinct and sometimes unique combination of physico-chemical parameters. One possible explanation of this observation is that receptors of one type tend to share a structurally conserved ligand-binding site. The structure of this site dictates the bundle of properties that a receptor-selective ligand should possess to properly bind the site. These properties include specific spatial, lipophilic, and H-binding parameters, as well as other features influencing pharmacodynamic behavior. On the other hand, the observed difference of physico-chemical properties for particular target-specific groups of peripheral fragments can result from different pharmacokinetic requirements for compounds acting on specific GPCR.

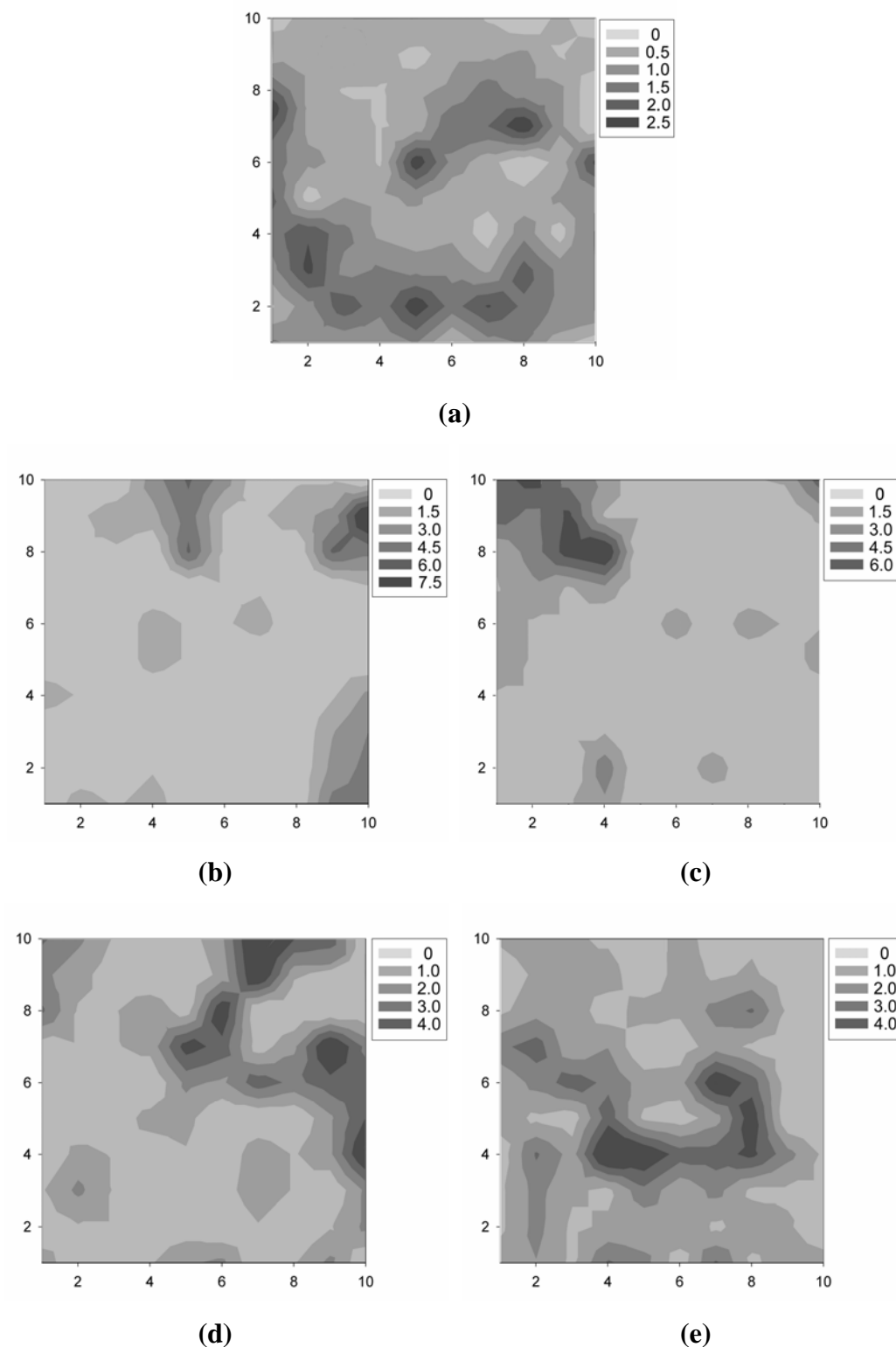


Figure 9. Distribution of five different target-specific groups of peripheral fragments within the Kohonen map: (a) Tachykinin NK1 antagonists (521 fragments), (b) Dopamine D2 agonists (113 fragments), (c) Cannabinoid CB1/CB2 agonists/antagonists (89 fragments), (d) β -3-Adrenoceptor agonists (294 fragments), (e) 5-HT_{1A} agonists (354 fragments). The data are in % (the total number of peripheral fragments in a receptor-specific group corresponds to 100%).

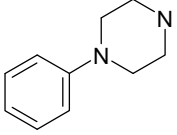
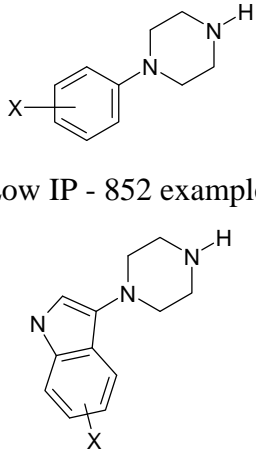
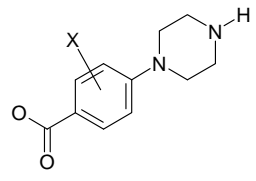
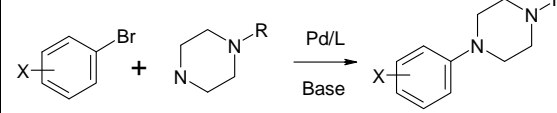
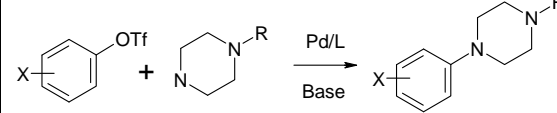
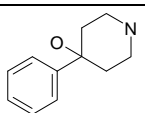
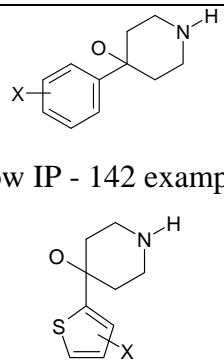
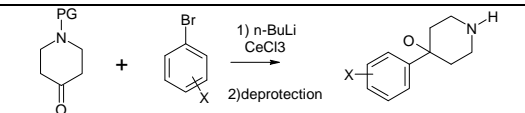
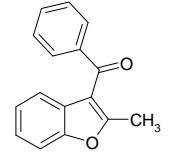
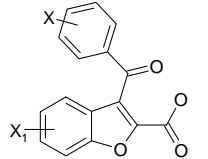
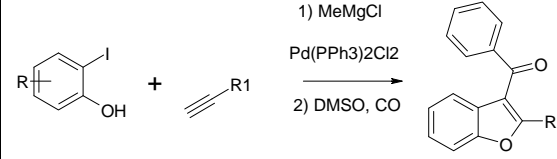
The observed differences create a basis for a rational selection of building blocks for synthesis of combinatorial libraries enriched in target-specific motifs. The quantitative structure-activity discrimination function found at this stage of our study can be used for effective search of reactive monomers possessing the desired physico-chemical and spatial parameters.

To summarize this part of our work, the privileged motifs can be considered, in a general case, as a main category of molecular fragments playing an essential role in the target-specific activity of a compound. On the other hand, the peripheral structural motifs are less important for target specificity and usually do not have any structural similarity, but, nevertheless, they are important for protein-ligand interactions and for a compound's pharmacokinetic profile. Modern chemical database management tools in combination with advanced methods of data mining permit effective identification of both privileged and peripheral molecular fragments. After these fragments are identified for each target activity group, they can be readily transformed into chemical building blocks for generation of a virtual target-biased combinatorial library.

3.7. Selection of Building Blocks

Combinatorial building blocks containing the privileged target-specific or nonselective structural motifs constitute the main category of reagents for synthesis. Based on statistical analysis of the total fragmented database, we created platforms of "privileged" core building blocks for all receptor-specific areas studied. For each privileged substructure, we selected a set of closely related compounds containing this privileged fragment (or its bioisosteric analog) and one or more "points of diversity" for introduction of peripheral building blocks. Examples of privileged structural motifs and the related core building blocks are shown in Table VI. A significant step in selection of core building blocks is the estimation of IP potential of the resulting compounds using the Beilstein database based on a number of known structures for reported active compounds containing the particular substructure. The next step in our selection procedure is related to assessment of synthetic accessibility (the third column in Table VI) and in generation of sets of assembling building blocks. It is important to note that such reagents sets, which can be used for synthesis of core building blocks, cannot be formed with the use of pseudo-retrosynthetic automatic approaches similar to RECAP. In most cases, the trivial simple cleavage rules used in such algorithms do not correspond to the practical methods of synthetic assembling of complex privileged structures (examples 3 and 4 in Table VI).

Table VI. Examples of core and assembling building blocks structurally related to privileged motifs belonging to different target-specific ligand groups

| Privileged substructures (unselective/target selective) | Core building blocks IP potential (Beilstein score) | Assembling building blocks. Example of reaction. |
|---|---|--|
|  <p>unselective* Dopamine D2 agonists, Tachykinin NK1 antagonists, α1 Adrenoceptor antagonists, PGE2 antagonists, 5-HT1D agonists, Vasopressin V1A/V2 antagonists, CCKB antagonists, 5-HT2A antagonists, Muscarinic M1 agonists, etc.</p> |  <p>Low IP - 852 examples</p>  <p>High IP - 3 examples</p> <p>High IP - 8 examples</p> |  <p>[xxv]</p>  <p>[xxvi]</p> |
|  <p>unselective Dopamine D2 antagonists, Tachykinin NK3 antagonists, Neurokinin NK3 antagonists, etc.</p> |  <p>Low IP - 142 examples</p> <p>High IP - 2 examples</p> |  <p>[xxvii]</p> |
|  <p>target specific Tachykinin NK1 antagonists</p> |  <p>High IP - 11 examples</p> |  <p>[xxviii]</p> |

* *N*-Arylpiperazines are typical privileged substructure in broad sense and have been used frequently in combinatorial synthesis^{xxix}. MDDR (August 2003 issue) contains 3449 physiologically active compounds containing *N*-arylpiperazine moiety, including 91 structures in clinical trials (55 – Phase I, 35 – Phase II, and 1 – Phase III), across more than 20 therapeutic areas

As it was discussed above, the peripheral building blocks determine pharmacokinetic and pharmacodynamic properties of compounds, as well as their target specificity in the case of compounds with unselective privileged cores. The selection of peripheral building blocks for the design of GPCR-targeted combinatorial libraries is based on application of Kohonen neural networks. It is important to note that before this experiment, all reagent structures should be reduced to their "normalized" representations to allow correct comparison with the structures of retrosynthetic fragments. For example, all carboxylic acid derivatives, such as acid chlorides, anhydrides or activated esters are transformed into their reduced radical form identical to that obtained after dissecting the amide bond; all alkyl halides and alcohols are transformed into alkyl radicals; etc. Such a "normalized" database of building blocks has been used in all neural network experiments. This database of available building blocks was processed on the same Kohonen map described in a previous section. For illustration, we show the results of the selection procedure for Cannabinoid CB1/CB2 receptor ligands (Figure 10). The hashed zone restricts the area of preferable peripheral fragments for Cannabinoid CB1/CB2 agonists/antagonists. A total of 10221 building blocks falling into the restricted area, were selected for synthesis planning.

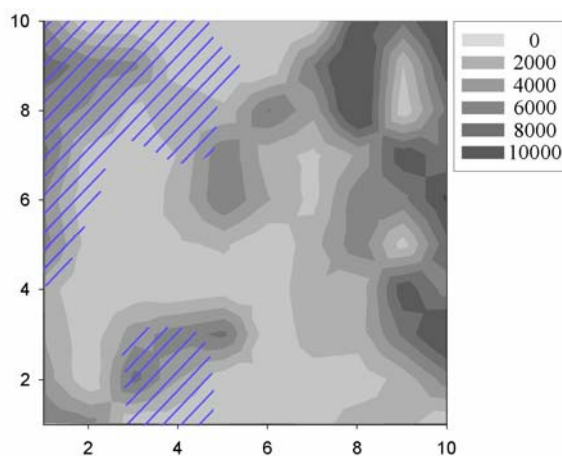


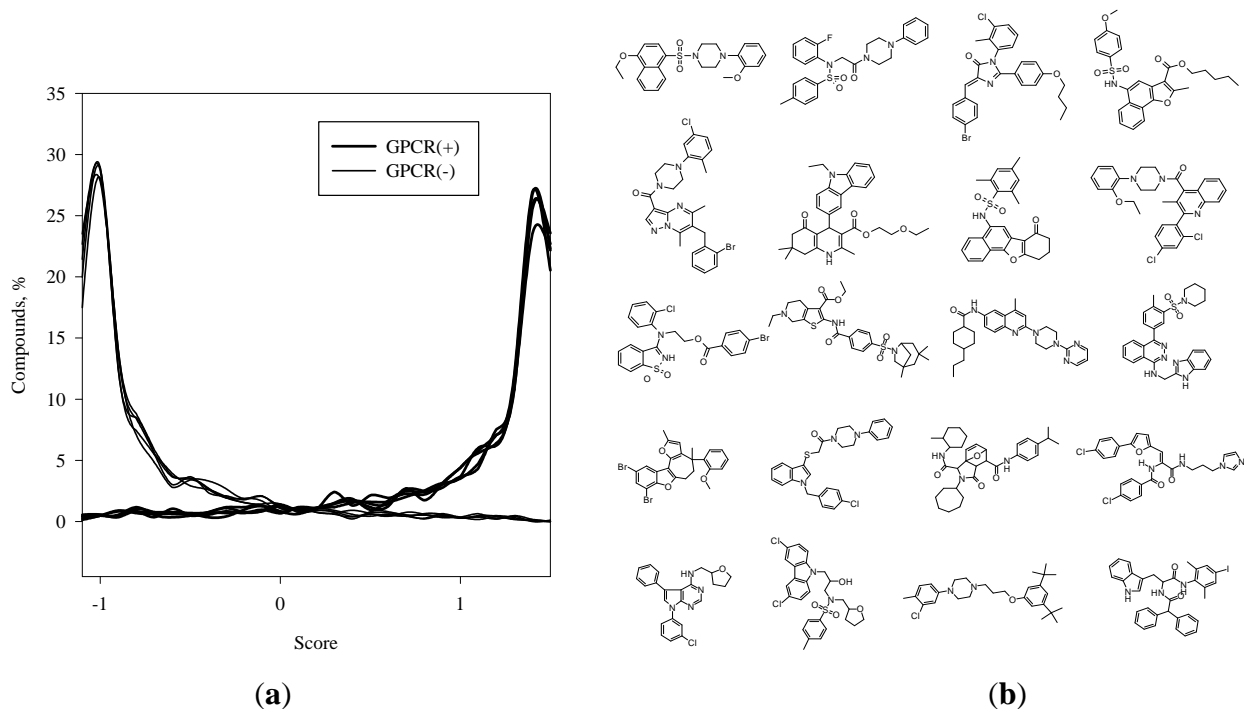
Figure 10. Distribution of reagents within the Kohonen map. The hatched zone restricts the area of preferable peripheral fragments for Cannabinoid CB1/CB2 agonists/antagonists

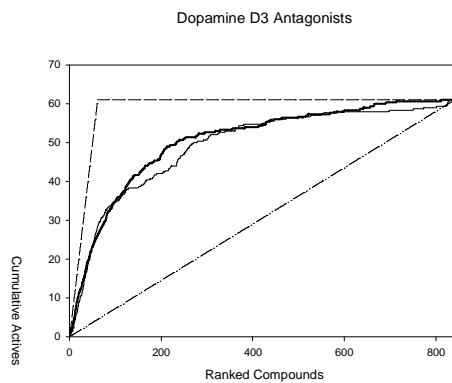
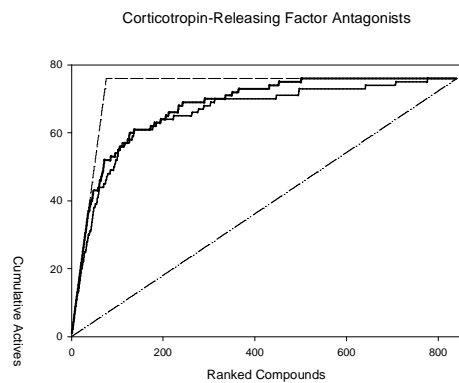
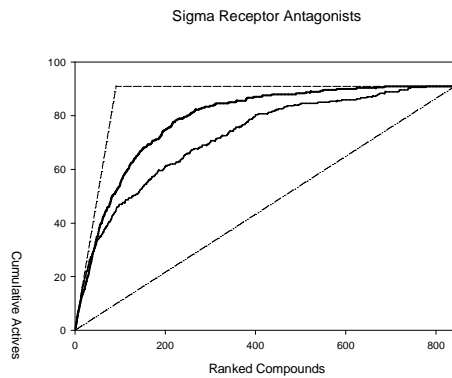
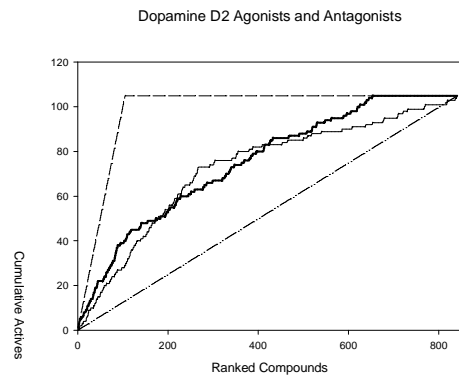
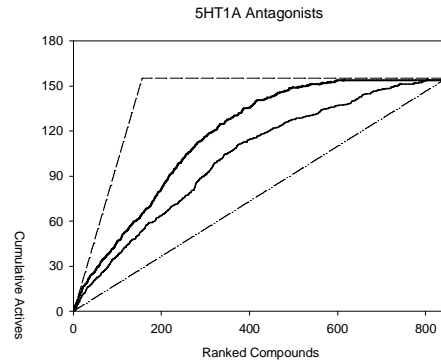
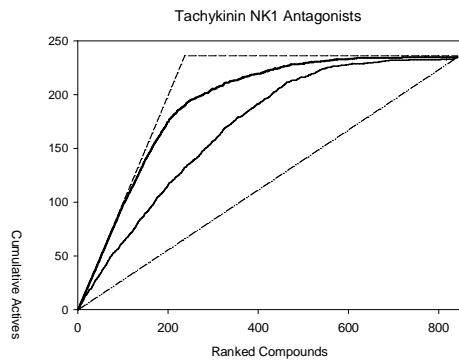
Selection of peripheral building blocks using the described method usually results in a relatively high number of candidates for synthesis. To reduce their number, additional, more stringent selection criteria can be applied. For instance, the structural similarity to the peripheral fragments found in the structures of active agents can be used. Optimization of structural diversity is another natural way to restrict the size of the initial selection. Additional filtering is related to exclude monomers that contain reactive chemical functions incompatible with the complementary functions of the privileged building

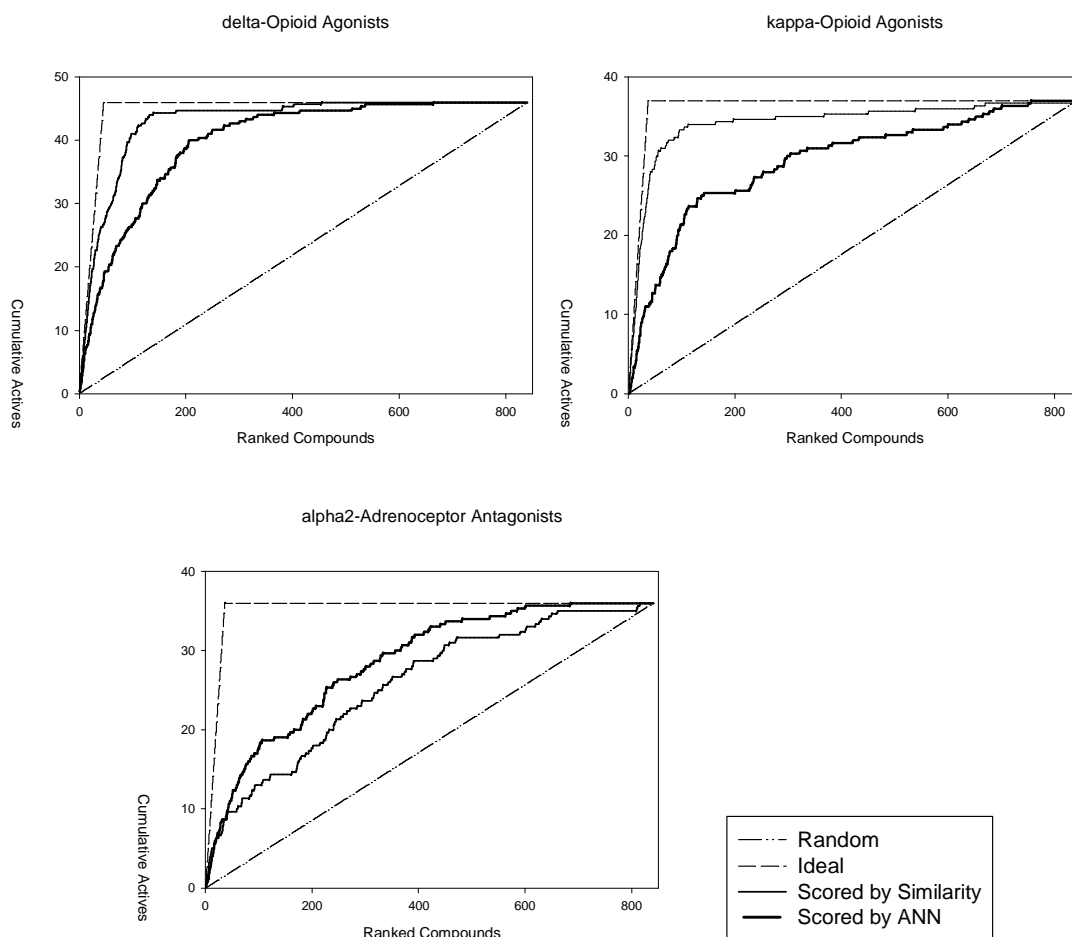
blocks. These algorithms and filtering procedures usually allow selection of 200-300 peripheral monomers for the generation of virtual combinatorial library targeted against a particular GPCR.

At the stage of virtual target-biased combinatorial library generation, the reagents containing the privileged structural motifs ("privileged building blocks") are categorized according to their reactive chemical functionality. Then the reagents with peripheral structural motifs ("peripheral building blocks") are divided into the chemical classes complementary to the corresponding privileged building blocks. After such a categorization is done, the automated procedure of virtual combinatorial library generation can be performed. Typically, it results in several tens of combinatorial libraries targeted against a particular GPCR containing a total of $10^4 - 10^6$ virtual compounds.

In addition, we have also used other computational techniques for GPCR-focused library design. These include Sammon mapping, SVM as well as NN modeling (for detail see^{xxx}). These methods are very effective for the more detailed analysis of the initially reduced set of compounds obtained after the second step of our virtual screening. The following Figure shows the core results obtained in our recent works. The models constructed have also been used for our GPCR-focused library design.







(c)

Figure 11. (a) The developed NN-model for GPCR library design (compound distributions on the scale of prediction scores for the test set; the data are shown for five independent randomizations)^{29a}; (b) representative examples of highest scoring structures selected from the GPCR-targeted library^{29a}; (c) graph showing accumulation of actives vs ranked for nine target-specific series studied in^{29b}. The solid line is assigned to the ANN ranking procedure and the dashed line is assigned to the fragment similarity-based ranking procedure

Conclusion

Primary bioscreening of large exploratory libraries of small molecules produced by combinatorial synthesis remains a key element of modern drug discovery. The problem of enhancement of bioscreening effectiveness necessitates more serious attention to the quality of screening compound libraries. In this context, advanced cheminformatics technologies, aiming at selection of the proper screening candidates, are of great industrial demand. The further evolution of such technologies will result in the development of integrated cheminformatics platforms, where all the issues related to selection of a rational pharmaceutically relevant screening candidate, having good synthetic feasibility, a desirable profile of target-specific action, drug-likeness, unexploited IP position, favorable ADME/Tox profile, compatibility with assay protocol, etc., will be solved with maximal quality, time- and cost-effectiveness.

The computational algorithm described here is very useful in constraining the size of virtual libraries of potential GPCR active agents. It can be effectively applied as an *in silico* filter to assist in the product-based design and planning of novel combinatorial libraries. Commonly, the described methodology can be generalized to aid in the selection of an optimal methodology for any arbitrary target-specific library design; it is not restricted to the GPCR-targeted libraries studied here. In addition, the results can be used for profiling the bioactivity of compounds based on comparison with the structures of known agents possessing a certain biological activity. The developed approach combines reagent- and product-based selection procedures, and results in generation of a compact virtual compound library (in the general case, several thousand compounds) targeted against a particular GPCR target. Usually such a library consists of several tens of distinct medium-sized combinatorial sub-libraries (50-200 compounds each) rich in target-specific structural motifs and possessing optimized physico-chemical properties. Such libraries represent a very useful tool at early stages of drug discovery and development, as a valuable source of primary hits easily amenable to further hit-to-lead optimization. Examples of particular GPCR-, protein kinase- and ion channel-targeted libraries generated using the described strategy can be found among the commercial products currently available at Chemical Diversity Labs, Inc.

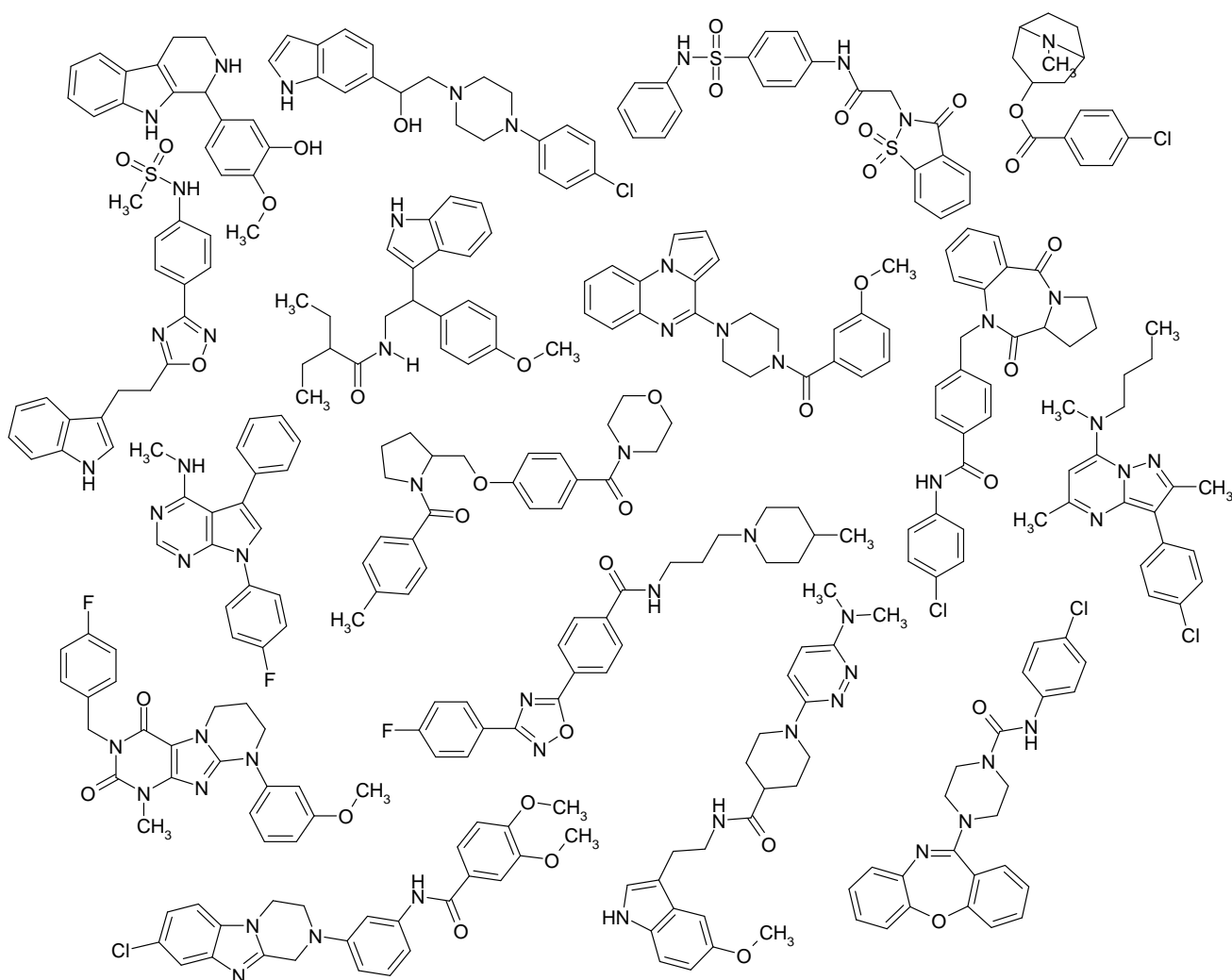


Figure 12. Representative structures from GPCR-targeted library

The described statistics-based method consists of a series of automated procedures, is easily reproducible, and can be recommended for practical design of compound libraries targeted against biotargets, for which sufficiently large number of selective ligands is available. It is important to note that all the described procedures are computationally inexpensive and permit real-time calculations with moderate hardware requirements. In particular, we have developed a Kohonen neural network-based model for *in silico* profiling of particular GPCR activity of small molecule drug-like compounds. It should be noted, that strategies for target-specific library design focused on structures of known ligands fail to adequately address the key issue of novelty of identification of novel active chemotypes. In practice, the ligand structure and property based approaches are used in combination with other methods, taking into account as much information as possible. For example, our NN methodology can be expended by selecting of similar substructures and bioisosteric analogs of known agents with specific action against particular GPCRs. This combined strategy is successfully applied at ChemDiv, Inc for the design of new generation of GPCR-targeted libraries enriched with novel lead chemotypes with significantly increased the hit rates. In our opinion, the increasing popularity of neural networks in

rational drug design is due to two main factors: the growing availability of quality structural data on ligands and targets, and the expanding computational power of modern computers. Thus, we have used the power of unsupervised neural network learning approach in the design of highly specific compound libraries targeted for several therapeutically significant GPCRs. This property-based approach is primarily applicable for the design of target-specific libraries enriched with novel ligand chemotypes. Certainly, these libraries can only be validated in primary screening.

The developed GPCR-related libraries are updated quarterly based on a “cache” principle. Older scaffolds/compounds are replaced by templates resulting from our in-house development (unique chemistry, literature data, computational approaches) while the overall size of the library remains the same (ca. 37K compounds). As a result, the libraries are renewed each year, proprietary compounds comprising 50-75% of the entire set. Clients are invited to participate in the template selection process prior to launch of our synthetic effort. Representative compounds from our GPCR-targeted library (a total of 37,032 compounds) are shown in Figure 12.

References

- ⁱ (a) T. Gudermann, B. Nurnberg, G. Schultz, *J. Mol. Med.* 1995, 73, 51 – 63; (b) H. Hamm, *J. Biol. Chem.* 1998, 273, 669 – 672.
- ⁱⁱ J. Drews, *Science* 2000, 287, 1960 – 1964.
- ⁱⁱⁱ J. Drews; *Science* 287 1960-1964 (2000).
- ^{iv} (a) Milligan G, Kostenis E. Heterotrimeric G-proteins: a short history. *Br J Pharmacol.* 2006 Jan;147 Suppl 1:S46-55; (b) Tenno T, Hiroaki H. Recent progress on structural biology of GPCR: perspectives and limits. *Tanpakushitsu Kakusan Koso.* 2008 Mar;53(3):256-64; (c) Römler H, Stäubert C, Thor D, Schulz A, Hofreiter M, Schöneberg T. G protein-coupled time travel: evolutionary aspects of GPCR research. *Mol Interv.* 2007 Feb;7(1):17-25.
- ^v (a) A. Ajay, W.P. Walters, M.A. Murcko; *J. Med. Chem.* 41 3314-3324 (1998); (b) A. Ajay; G.W. Bemis, M.A. Murcko; *J. Med. Chem.* 42 4942-4951 (1999); (c) J. Sadowski, H.A. Kubinyi; *J. Med. Chem.* 41 3325-3329 (1998).
- ^{vi} (a) K.V. Balakin, S.E. Tkachenko, S.A. Lang, I. Okun, A.I. Ivashchenko, N.P. Savchuk; Property-Based Design of GPCR-Targeted Library *J. Chem. Inf. Comput. Sci.* 42 1332-1342 (2002); (b) Konstantin V. Balakin*, Stanley A. Lang, Andrey V. Skorenko, Sergey E. Tkachenko, Andrey A. Ivashchenko, Nikolay P. Savchuk. *JCICS. Structure-Based versus Property-Based Approaches in the Design of G-Protein-Coupled Receptor-Targeted Libraries*; (c) Ivanenkov Y.A., Balakin K.V., Skorenko A.V., Tkachenko S.E., Savchuk N.P., Ivachtchenko A.A., Nikolsky Y. Application of advanced machine learning algorithm for profiling specific GPCR-active compounds. *Chem. Today.*, 2003, 21, 72-75.
- ^{vii} N.P. Savchuk, S.E. Tkachenko, K.V. Balakin. Rational Design of GPCR-specific Combinational Libraries Based on the Concept of Privileged Substructures / In *Chemoinformatics in Drug Discovery*, Ed. Prof. Dr. Tudor I. Oprea, Wiley-VCH Verlag GmbH & Co. KGaA, 2005.
- ^{viii} Nikolay P. Savchuk, Sergey E. Tkachenko, Konstantin V. Balakin. *Strategies for the Design of pGPCR Targeted Libraries*, Wiley, 2006, VOL 30, pages 137-164.
- ^{ix} H. van de Waterbeemd, D.A. Smith, K. Beaumont, D.K. Walker; *J. Med. Chem.* 44 1313-1333 (2001).

- ^x S. Anzali, J. Gasteiger, U. Holzgrabe, J. Polanski, J. Sadowski, A. Teckentrup, M. Wagener; "The Use of Self-Organizing Neural Networks in Drug Design" in 3D QSAR in Drug Design – Vol.2, H. Kubinyi, G. Folkers, Y. C. Martin, Eds.; Kluwer/ESCOM, Dordrecht, NL, 1998, pp.273-299.
- ^{xi} H. Bauknecht, A. Zell, H. Bayer, P. Levi, M. Wagener, J. Sadowski, J. Gasteiger; *J. Chem. Inf. Comp. Sci.* 36 1205-1213 (1996)
- ^{xii} S. Anzali, G. Barnickel, M. Krug, J. Sadowski, M. Wagener, J. Gasteiger, J. Polanski; *J. Comp.-Aid. Mol. Des.* 10 521-534 (1996).
- ^{xiii} M. Brüstle, B. Beck, T. Schindler, W. King, T. Mitchell, T. Clark; *J. Med. Chem.* 45 3345-3355 (2002).
- ^{xiv} W. P. Walters, M. T. Stahl, M. A. Murcko, *Drug Disc. Today* **1998**, 3, 160 – 178.
- ^{xv} D. E. Clark, S. D. Pickett, *Drug Disc. Today* **2000**, 5, 49 – 58.
- ^{xvi} P. J. Eddershaw, A. P. Beresford, M. K. Bayliss, *Drug Disc. Today* **2000**, 5, 409 – 414.
- ^{xvii} M. J. Valler, D. Green, *Drug Disc. Today* **2000**, 5, 286–293.
- ^{xviii} Prous Science Integrity Database: <http://integrity.prous.com/integrity/servlet/xmlxsl>
- ^{xix} S. V. Trepalin, V. A. Gerasimenko, A. V. Kozyukov, N. Ph. Savchuk, A. A. Ivashchenko, New Diversity Calculations Algorithms Used for Compound Selection. *J. Chem. Inf. Comput. Sci.* **2002**, 42, 249 – 258.
- ^{xx} www.chemdiv.com
- ^{xxi} (a) J. Gasteiger, M. Marsili, M. G. Hutchings, H. Saller, P. Loew, P. Roese, K. Rafeiner, *J. Chem. Inf. Comp. Sci.* 1990, 30, 467 – 476; (b) J. E. Ridings, M. D. Barratt, R. Cary, C. G. Earnshaw, C. E. Eggington, M. K. Ellis, P. N. Judson, J. J. Langowski, C. A. Marchant, *Toxicology* 1996, 106, 267 – 279; (c) X. Q. Lewell, D. B. Judd, S. P. Watson, M. M. Hann, *J. Chem. Inf. Comp. Sci.* 1998, 38, 511 – 522; (c) G. Schneider, O. Clément-Chomienne, L. Hilfiger, P. Schneider, S. Kirsch, H. Böhm, W. Neidhart, *Angew. Chem. Int. Ed.* 2000, 39, 4130 – 4133.
- ^{xxii} X. Q. Lewell, D. B. Judd, S. P. Watson, M. M. Hann, *J. Chem. Inf. Comp. Sci.* 1998, 38, 511 – 522.
- ^{xxiii} G. A. Patani, E. J. LaVoie, *Chem. Rev.* **1996**, 96, 3147 – 3176.
- ^{xxiv} D. Korolev, K. V. Balakin, Y. Nikolsky, E. Kirillov, Y. A. Ivanenkov, N. P. Savchuk, A. A. Ivashchenko, T. Nikolskaya. *J. Med. Chem.* **2003**, 46, 3631 – 3643.
- ^{xxv} U. Singh, E. Strieter, D. Blackmond, S. Buchwald, *J Am Chem Soc.* **2002**, 124, 14104 – 14114.
- ^{xxvi} J. Wolfe, S. Buchwald, *J. Org. Chem.* **1997**, 62, 1264 – 1267.
- ^{xxvii} J. Albert, D. Aharony, D. Andisik, H. Barthlow, H. Bernstein, R. Bialecki, R. Dedinas, R. Dembofsky, D. Hill, K. Kirkland, G. Koether, B. Kosmider, C. Ohnmacht, W. Palmer, W. Potts, W. Rumsey, L. Shen, L. Shenvi, S. Sherwood, S. Warwick, K. Russell, *J. Med. Chem.* **2002**, 45, 3972 – 3983.
- ^{xxviii} R.V.A. Orru, de Greef, *Synthesis.* **2003**, 10, 1471 – 1499.
- ^{xxix} P. Baraldi, M. Nunez, A. Morelli S. Falzoni, F. Di Virgilio, R. Romagnoli, *J Med Chem.* **2003**, 46, 1318 – 1329.
- ^{xxx} (a) Konstantin V. Balakin, Sergey E. Tkachenko, Stanley A. Lang, Ilya Okun, Andrey A. Ivashchenko, and Nikolay P. Savchuk. Property-Based Design of GPCR-Targeted Library. *J Chem Inf Comput Sci.* 2002 Nov-Dec;42(6):1332-42; (b) Konstantin V. Balakin*, Stanley A. Lang, Andrey V. Skorenko, Sergey E. Tkachenko, Andrey A. Ivashchenko, Nikolay P. Savchuk. Structure-Based versus Property-Based Approaches in the Design of G-Protein-Coupled Receptor-Targeted Libraries. *J. Chem. Inf. Comput. Sci.*, 2003, 43 (5), pp 1553–1562.