# Design and Exploration of Target-Selective Chemical Space Representations

Ingo Vogt and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology & Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstrasse 2, D-53113 Bonn, Germany

Received March 27, 2008

We report the design of target-selective chemical spaces using CA-DynaMAD, a mapping algorithm that generates and navigates flexible space representations for the identification of active or selective compounds. The algorithm iteratively increases the dimensionality of reference spaces in a controlled manner by evaluating a single descriptor per iteration. For seven sets of closely related biogenic amine G protein coupled receptor (GPCR) antagonists with different selectivity, target-selective reference spaces were designed and used to identify selective compounds by screening a biologically annotated database. Combinations of descriptors that constitute target-selective reference spaces identified with CA-DynaMAD can also be used to build other computational models for the prediction of compound selectivity.

## INTRODUCTION

The generation of chemical reference spaces is an integral component of many chemoinformatics approaches including the majority of molecular similarity-based methods.[1–3] Many compound classification[2,3] and library design[4,5] methods utilize arrays of molecular property descriptors[6,7] to construct feature spaces that provide a basis for the analysis of molecular similarity relationships or compound diversity distributions. For the evaluation and prediction of biological activity in the context of, for example, target-focused library design[5] or virtual compound screening[3,8] the relevance of the chosen chemical space representations for targeted activities is of paramount importance.[8,9] Simply put, biological activities cannot be deduced or predicted from molecular structure and properties if selected descriptors are not responsive to activity-determining features. Thus, descriptor selection is a crucially important task for the study of structure/property-activity relationships in multidimensional chemical space representations. The importance of feature selection in chemical space design has long been recognized, and a number of attempts have been made to rationalize this process and ensure that chemical reference spaces are appropriate for the problems under investigation. For example, the "receptor-relevant subspace" concept[10] attempts to study compounds in reference spaces formed by complex orthogonal descriptors that combine chemical features generally known to be important for mediating specific receptor−ligand interactions. Such space representations can also be employed for the design of target-focused libraries.[5] Underlying ideas include that so-generated reference spaces are generally relevant for the study of receptor−ligand interactions and that compounds that preferentially populate certain subspaces and cluster along selected descriptor axes are likely to share similar biological activities.[10] Another generally applicable approach to the design of chemical reference spaces is the selection of those descriptors from high-dimensional space representations that are most important to account for the feature variance within a set of diverse or active compounds and construct lower-dimensional space representations from them.[11] Furthermore, chemical space design can also be attempted by systematically searching for descriptor combinations that group classes of known active compounds together in chemical space and distinguish them from others.[12] In addition, concepts from information theory have been applied to identify descriptors that have high information content in compound classes under study or differences in information content between, for example, active and database molecules.[13]

Essentially, the above-mentioned approaches have in common that they attempt to generate low-dimensional space representations that are often thought to be preferred for compound classification or the analysis of compound distributions.[11,14] Only recently, methods have been developed or adapted that deliberately depart from this paradigm of low-dimensionality. These approaches navigate high-dimensional space representations that are tailored toward selected compound classes and include support vector machines[15–18] for binary compound classification (class label prediction) and, in addition, a class of methods termed mapping algorithms. The latter methods attempt to facilitate activity class-directed chemical space design by selecting combinations of molecular descriptors that have systematically different settings or value ranges in different compound sets (e.g., activity classes and database compounds). Thus far, three mapping algorithms have been introduced including Dynamic Mapping of Consensus positions (DMC),[19] Mapping to Activity class-specific Descriptor value ranges (MAD),[20] and Dynamic MAD (DynaMAD).[21] The MAD algorithm introduced a static mapping procedure of compounds to activity class-selective descriptor value ranges that were determined using a statistical scoring function. This scoring function calculated the probability of an arbitrary database compound to match the descriptor value range of an activity class. For MAD, a preselected number of descriptors with activity class-selective tendencies were used

* Corresponding author phone: +49-228-2699-306; fax: +49-228-2699-341; e-mail: bajorath@bit.uni-bonn.de.

10.1021/ci800106e CCC: $40.75 
Published on Web 06/18/2008

for mapping of database compounds. In DynaMAD, dynamic compound mapping was implemented. This type of dynamic compound mapping was first introduced with DMC, which utilized simplified (binary-transformed) descriptor spaces. This makes DynaMAD the conceptually most advanced of these methods because it operates in original unmodified descriptor spaces of stepwise increasing dimensionality. Dynamic mapping iteratively adds descriptor subsets of decreasing selectivity until the majority of database compounds are eliminated from chemical space representations focusing on active reference molecules. Thus, for dimension extension, descriptors are organized in different layers according to precalculated activity class-specific tendencies,[20] and several descriptors (i.e., dimensions) are added to the evolving chemical space representations per dimension extension step. Dimension extension is continued until the chemical reference space is highly resolved and capable of distinguishing the vast majority of database decoys from compounds having the desired activity.

In extensive benchmark calculations, DynaMAD has been shown to produce significant recall of active molecules when searching databases containing millions of compounds.[21] Recently, DynaMAD has also been applied to analyze compound selectivity.[22] In this study, different computational methods were evaluated to distinguish between compounds having different selectivity against closely related targets. Both DynaMAD and state-of-the-art fingerprints displayed a tendency to detect target-selective compounds in databases when selective reference molecules were used.[22] These findings suggested that selectivity differences between molecules that are active against related targets might be predictable using currently available similarity-based methods.

Given the ability of DynaMAD to generate compound class-directed chemical space representations, we have asked the question whether it might be possible to systematically design target-selective chemical spaces. Compared to searching for active compounds, assessing compound selectivity is further complicated because it is required to distinguish chemically related compounds from each other that have differential activities against multiple members of a target family. Consequently, chemical reference spaces for such tasks must be designed in previously unexplored ways. However, for applications in chemical biology and medicinal chemistry, computational methods that are capable of analyzing and predicting ligand selectivity profiles within target families are highly attractive.[23] Therefore, in order to investigate the design of target-selective chemical spaces, we have developed a second generation DynaMAD-like mapping algorithm termed Continuous-Adaptive DynaMAD (CA-DynaMAD) that further refines feature selection and better controls dimension extension. The algorithm evaluates one descriptor per step and performs descriptor value range analysis on the basis of the continuously updated (size-reduced) background compound database. Using this approach, we have successfully generated target-selective descriptor spaces for seven sets of closely related GPCR antagonists. Each space representation distinguished one compound set from all others. The dimensionality of these spaces was further increased using the CA-DynaMAD dimension extension function in order to distinguish large numbers of database compounds from selectivity sets and identify target-selective antagonists in a biologically annotated database.

**Table 1.** Composition of Selectivity Sets[a]

| GPCR | selective over | no. of compounds | selectivity range |
|---|---|---|---|
| 5HT1a | 5HT2a, Alpha1, D2 | 53 | 65−476190 |
| 5HT2a | 5HT1a, Alpha1, D1, D2 | 21 | 67−49122 |
| Alpha1 | 5HT1a, 5HT2a, D1, D2, D3, D4 | 26 | 59−62500 |
| D1 | D2, D4 | 33 | 55−10084 |
| D2 | 5HT1a, D1, D3, D4 | 25 | 73−18310 |
| D3 | D1, D2, D4 | 37 | 51−17600 |
| D4 | Alpha1, D1, D2, D3 | 72 | 54−28000 |

[a] Selectivity ranges were determined by calculating the ratio of $K_i$ values for all pairwise comparisons of compounds with reported activity against the listed receptors.[22] "D" abbreviates dopamine.

## MATERIALS AND METHODS

**Compound Selectivity Sets and Descriptors.** The computational study of molecular selectivity critically depends on the availability of appropriate compound sets that contain molecules with experimentally determined binding profiles against multiple targets. In our analysis, we have used a previously designed system consisting of selectivity sets for seven closely related biogenic amine GPCRs and containing a total of 267 compounds.[22] The composition of these selectivity sets is summarized in Table 1. Selective compounds were assembled from original literature sources.[22] A selectivity threshold of at least 50-fold higher potency for one receptor over at least one other was required to assign a compound to a selectivity set. Figure 1 shows representative compounds from each set. Each selectivity set contains different scaffolds, and the compounds in these seven sets represent a structural spectrum with in part significant interset similarity.[22]

As a descriptor basis set for our analysis, 155 1D and 2D molecular property descriptors available in the Molecular Operating Environment[24] were used.

**CA-DynaMAD.** DynaMAD was originally developed for generating and navigating high-dimensional chemical space representations and efficient processing of very large compound databases.[21] The method automatically selects descriptors from basis sets using a descriptor scoring function that calculates the probability of a database compound to map the descriptor value range of a set of reference molecules (typically a set of active compounds). If the probability of a database compound to fall into the reference set value range of a descriptor is low, then the descriptor displays a reference set-specific tendency. Thus, descriptors are scored and rank-ordered according to compound class specificity and assigned to different dimension extension levels based on their scoring range. During dimension extension, all descriptors of each dimension extension level are added in subsequent steps, and database compounds are mapped to the resulting reference set-specific consensus positions. Only those compounds are retained that match all descriptor reference set value ranges; others are discarded. This process is continued until only a small compound selection set remains. As a termination criterion, the maximum number of remaining database compounds is typically set to 100 or fewer.

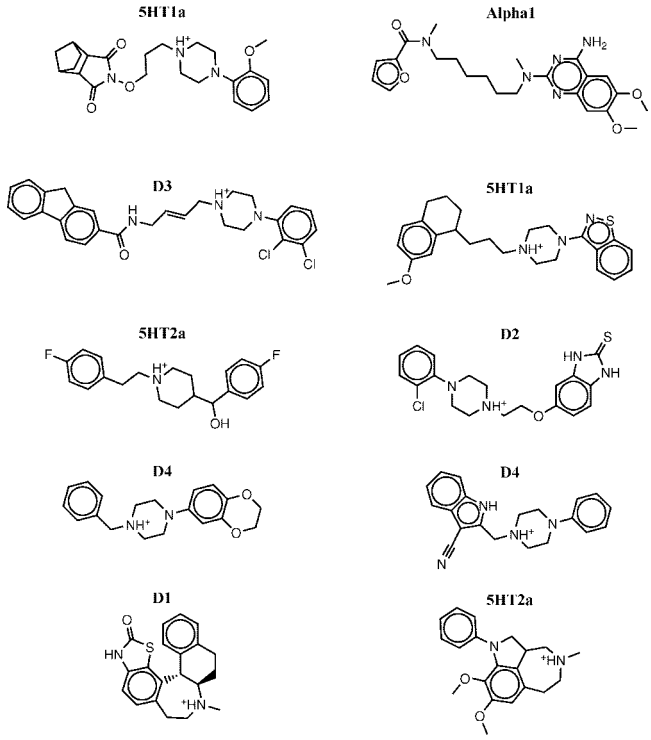


**Figure 1.** Exemplary structures of selective GPCR antagonists. At the top left and right, the most selective 5HT1a and Alpha1 antagonists are shown, respectively. In addition, antagonists from other selectivity sets with varying degrees of similarity to these highly selective compounds are shown. The representation illustrates a part of the structural spectrum of GPCR antagonists with different selectivity.

In order to transform dimension extension into a continuous, rather than a stepwise, function and better control the selection of most relevant descriptors, the algorithm was modified in two ways. First, descriptors are added one-by-one in the order of decreasing scores (corresponding to a continuous increase in reference space dimensionality). If several descriptors produce the same score, the descriptor with lowest average correlation to all previously selected ones is chosen, and compound mapping is carried out. Then the process continues. Second, adaptive descriptor scoring is introduced, which is explained as follows. Scoring requires the comparison of reference set value ranges and database distributions of descriptors. In the original implementation of DynaMAD, descriptor distributions are only calculated once for the background database that typically contains very large numbers (i.e., millions) of compounds, and the descriptors are scored on the basis of these distributions. However, during dimension extension, the background database is continuously reduced in size, which is changing descriptor value distributions, in particular, when large numbers of database compounds are eliminated. Therefore, adaptive scoring repeats the descriptor scoring process at each step with recalculated distributions for the size-reduced database and the descriptor ranking is updated. These continuous descriptor evaluation and adaptive descriptor scoring schemes are implemented in CA-DynaMAD.

**Benchmark Calculations.** In order to compare the performance of DynaMAD and CA-DynaMAD, systematic test calculations were carried out. Each selectivity set was 100 times randomly divided, and 50% of the selective compounds were used as a reference set. The remaining 50% were added to an in-house generated 2D-unique version of the ZINC database[25] (~3.7 million compounds) as potential selective hits. In total, 100 independent search calculations were carried out for each selectivity set, and each calculation was terminated when 100 or fewer database compounds remained.

**Flexible Design of Target-Selective Chemical Spaces.** In order to identify descriptor combinations that distinguish a selectivity set from all others, each selectivity set was used once as a reference set and the remaining six sets as test compounds. CA-DynaMAD calculations were carried out until all, or almost all, of these GPCR test compounds were eliminated through addition of high-scoring reference set descriptors. The resulting descriptor combinations form target-selective chemical reference spaces.

The MDL Drug Data Report (MDDR)[26] containing ~160,000 compounds was then mapped to the target-selective descriptor spaces, and the number of MDDR compounds matching all selectivity set value ranges was determined. For each selectivity set, dimension extension was continued until 100 or 50 molecules remained. The "activity" and "action" fields in the entries of all compounds in MDDR selection sets were inspected to determine whether antagonists and target-selective hits were identified. Prior to this analysis, selectivity set compounds also present in the MDDR were removed from the database.

## RESULTS AND DISCUSSION

**CA-DynaMAD versus DynaMAD.** In order to compare CA-DynaMAD and DynaMAD, test calculations were carried out on our seven compound selectivity sets. Akin to typical virtual screening benchmark calculations, subsets of target-selective compounds were used as reference sets to search for the remaining selective molecules added to a background database. For each selectivity set, a total of 100 independent search calculations were carried with varying reference set composition. Thus, a total of 1400 search calculations were performed. In 1376 of these trials, final database selection sets of 100 or fewer compounds contained correctly identified target-selective molecules. DynaMAD calculations failed to recover selective compounds in 23 individual trials (four, six, and 13 for Alpha1, 5HT2a, and D2, respectively), whereas CA-DynaMAD failed only in a single of its 700 trials (for Alpha1). A comparison of average results is shown in Figure 2. The effects of using a continuous dimension extension and adaptive descriptor scoring on the dimensionality of the final space representations can be estimated in Figure 2a. For compound mapping using CA-DynaMAD, significantly fewer descriptors were required than for DynaMAD, on average, only about a third. Figure 2b illustrates that hit rates achieved with CA-DynaMAD and DynaMAD were overall comparable, despite in part significant differences in the number of descriptors that were required. However, as shown in Figure 2c, the recovery rates of selective compounds using CA-DynaMAD were consistently higher. On average, about 25% more target-selective compounds were recovered with CA-DynaMAD than with DynaMAD. Continuous dimension extension better controls the gradual reduction of database size, and adaptive descriptor scoring reduces the number of potential hits that are lost during dimension extension. This is the case because

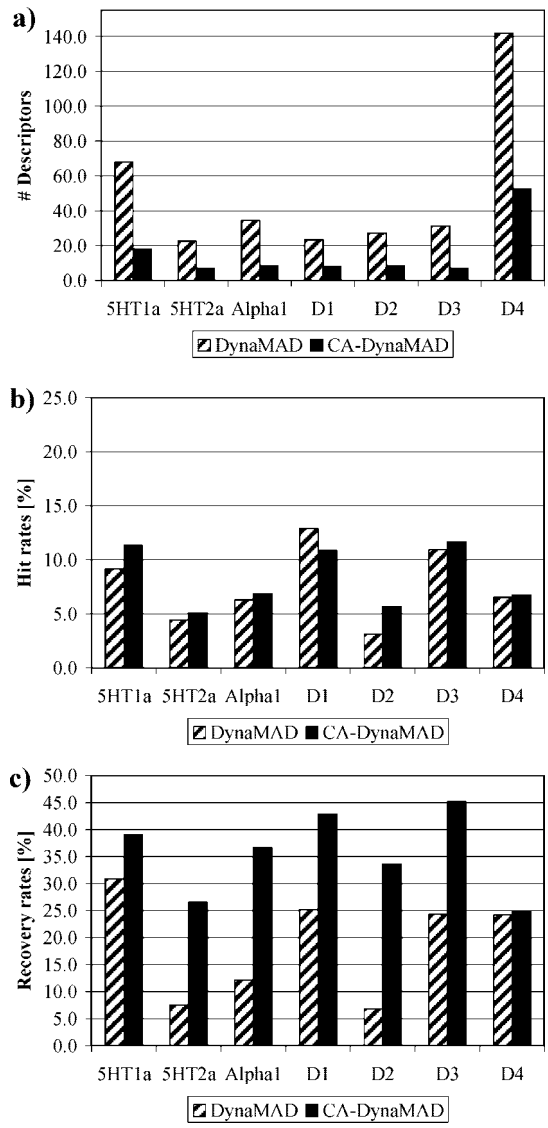


**Figure 2.** Virtual screening trials. Average results of 100 search calculations are reported for each selectivity set. In **a)**, the number of descriptors for DynaMAD and CA-DynaMAD calculations is given; **b)** and **c)** report hit and recovery rates at the final dimension extension steps, respectively.

descriptors that discriminate most effectively between reference and database compounds are redetermined at each step. Thus, the test calculations confirmed our expectations and the desired improvements of CA-DynaMAD over the original DynaMAD implementation.

**Target-Selective Descriptor Spaces.** Using CA-DynaMAD, we next investigated the design of target-selective space representations. In these calculations, compound mapping was carried out using each complete selectivity set as a reference until all compounds or the maximum number of compounds belonging to the other selectivity sets were eliminated. Ideally, the termination criterion would be elimination of all other compounds. However, depending on the degree of selectivity of the reference space representation, this might not be possible in all cases. For five of seven selectivity sets, all other GPCR antagonists were eliminated, thus producing "pure" target-selective reference spaces. In the case of D4, a single 5HT1a compound remained and *vice versa*. Thus, even for selectivity sets D4 and 5HT1a, impurities were minute. Interestingly, only 29 of the 155 descriptors in our basis set contributed to target-selective space representations. A summary of these descriptors is provided in Table 2. Preferred descriptors included complex and orthogonal (i.e., information-rich) designs of the BCUT[14] and VSA[27] types, a number of topological indices, and even simple descriptors such as the fraction of rotatable bonds or the number of double bonds in a molecule. Of these 29 descriptors, 23 only contributed to one of seven target-selective space representations and four to two. The balabanJ index and the SlogP_VSA4 descriptor occurred three and four times, respectively. Thus, only a small number of property descriptors was required to successfully differentiate between GPCR antagonists having different selectivity, and the majority of these descriptors uniquely contributed to different reference spaces.

Table 3 reports the CA-DynaMAD mapping statistics for each of the seven selectivity sets. As can be seen, only three to eight descriptors were required to build target-selective space representations. During dimension extension, GPCR

**Table 2.** Descriptors for Target-Selective Chemical Spaces[a]

| type | descriptor | explanation |
|---|---|---|
| adjacency and distance matrix descriptors | balabanJ | Balaban's topological connectivity index[28] |
| | BCUT_PEOE_0<br>BCUT_PEOE_3<br>BCUT_SLOGP_1<br>BCUT_SLOGP_3<br>BCUT_SMR_0 | BCUT-type descriptors[14] for PEOE partial charges,[29] SlogP,[30] or molar refractivity (SMR)[30] |
| | GCUT_PEOE_0<br>GCUT_SMR_0<br>GCUT_SLOGP_0 | GCUT descriptors use graph distances instead of bond order information (like BCUTs) |
| | VDistEq | distance matrix index |
| atom and bond counts | b_1rotR | fraction of rotatable bonds |
| | b_double | number of double bonds |
| | lip_don | number of OH and NH groups |
| connectivity and shape indices | chi1 | connectivity index[31] |
| | KierA3 | shape index[31] |
| partial charge descriptors | PEOE_VSA+0<br>PEOE_VSA+1 | fractional polar van der Waals surface area (vdW_sa) |
| | PEOE_VSA_FPPOS | fractional positive (vdW_sa) |
| | PEOE_VSA_HYD | total hydrophobic (vdW_sa) |
| molecular surface descriptors | vsa_acc | approximate sum of vdW surface areas of hydrogen bond acceptors |
| | vsa_pol | approximate sum of vdW surface areas of polar atoms |
| subdivided surface area descriptors[27] | SlogP_VSA1<br>SlogP_VSA4<br>SlogP_VSA5<br>SlogP_VSA7 | combined with SlogP[30] |
| | SMR_VSA1<br>SMR_VSA2<br>SMR_VSA4<br>SMR_VSA6 | combined with molar refractivity (SMR[30]) |

[a] Descriptors are abbreviated and explained according to the Molecular Operating Environment and its documentation material.

**Table 3.** Generation of Target-Selective Chemical Space Representations[a]

| class | step | descriptor | deselection [%] | purity [%] |
|---|---|---|---|---|
| | | a) | | |
| 5HT1a | 1 | chi1 | 54.7 | 35.3 |
| | 2 | balabanJ | 20.6 | 50.0 |
| | 3 | SMR_VSA1 | 11.2 | 64.6 |
| | 4 | SlogP_VSA1 | 6.1 | 76.8 |
| | 5 | PEOE_VSA+0 | 3.3 | 85.5 |
| | 6 | GCUT_PEOE_0 | 1.4 | 89.8 |
| | 7 | VDistEq | 1.4 | 94.6 |
| | 8 | SlogP_VSA4 | 0.9 | 98.2 |
| | | b) | | |
| 5HT2a | 1 | BCUT_SLOP_3 | 76.4 | 26.6 |
| | 2 | PEOE_VSA+1 | 11.8 | 42.0 |
| | 3 | SlogP_VSA7 | 5.3 | 56.8 |
| | 4 | BCUT_SLOGP_1 | 3.3 | 72.4 |
| | 5 | GCUT_SMR_0 | 2.9 | 95.5 |
| | 6 | SMR_VSA2 | 0.4 | 100.0 |
| | | c) | | |
| Alpha1 | 1 | vsa_acc | 85.5 | 42.6 |
| | 2 | SMR_VSA4 | 12.4 | 83.9 |
| | 3 | b_1rotR | 1.7 | 96.3 |
| | 4 | BCUT_PEOE_3 | 0.4 | 100.0 |
| | | d) | | |
| D1 | 1 | BCUT_SMR_0 | 74.8 | 35.9 |
| | 2 | SlogP_VSA5 | 17.5 | 64.7 |
| | 3 | lip_don | 4.3 | 80.5 |
| | 4 | SMR_VSA6 | 2.6 | 94.3 |
| | 5 | balabanJ | 0.4 | 97.1 |
| | 6 | SlogP_VSA4 | 0.4 | 100.0 |
| | | e) | | |
| D2 | 1 | BCUT_PEOE_0 | 91.3 | 54.4 |
| | 2 | b_double | 6.2 | 80.7 |
| | 3 | vsa_pol | 2.5 | 100.0 |
| | | f) | | |
| D3 | 1 | KierA3 | 93.5 | 71.2 |
| | 2 | GCUT_SLOGP_0 | 4.8 | 90.2 |
| | 3 | SlogP_VSA4 | 1.7 | 100.0 |
| | | g) | | |
| D4 | 1 | PEOE_VSA_HYD | 50.8 | 42.9 |
| | 2 | BCUT_SLOGP_3 | 22.6 | 58.1 |
| | 3 | GCUT_SMR_0 | 14.4 | 75.0 |
| | 4 | balabanJ | 6.7 | 86.8 |
| | 5 | b_1rotR | 3.1 | 93.6 |
| | 6 | SMR_VSA4 | 1.0 | 96.0 |
| | 7 | SlogP_VSA4 | 0.5 | 97.3 |
| | 8 | PEOE_VSA_FFPOS | 0.5 | 98.6 |

[a] "Deselection" reports the percentage of the total number of GPCR antagonists with different selectivity that were eliminated through the addition of each descriptor. "Purity" gives the cumulative percentage of target-selective antagonists among all compounds passing the dimension extension step.

antagonists belonging to other selectivity sets were gradually eliminated. However, the top-scoring one or two descriptors made the most significant contributions. In five cases, more than 70% of other GPCR antagonists did not match the selectivity reference set value ranges of the first descriptors. Thus, the respective descriptor value ranges were already signatures of GPCR antagonist selectivity. For selectivity sets D2 and D3, the top-scoring descriptors (BCUT_PEOE_0 and KierA3, respectively) deselected more than 90% of the other antagonists. Taken together, these findings illustrate the ability of CA-DynaMAD calculations to identify preferred descriptors for chemical space design and compound classification.

**Table 4.** MDDR Compounds in Low-Dimensional Target-Selective Descriptor Spaces

| class | no. of descriptors | no. of MDDR compounds |
|---|---|---|
| 5HT1a | 8 | 1096 |
| 5HT2a | 6 | 1397 |
| Alpha1 | 4 | 7418 |
| D1 | 6 | 2339 |
| D2 | 3 | 2651 |
| D3 | 3 | 4153 |
| D4 | 8 | 1914 |

**Table 5.** High-Dimensional Target-Selective Spaces[a]

| class | dimensionality | no. of MDDR compounds |
|---|---|---|
| 5HT1a | 30 (+22) | 49 |
| 5HT2a | 14 (+8) | 50 |
| Alpha1 | 12 (+8) | 45 |
| D1 | 11 (+5) | 47 |
| D2 | 9 (+6) | 44 |
| D3 | 10 (+7) | 48 |
| D4 | 59 (+51) | 113 |

[a] The numbers in parentheses report the increase in dimensionality relative to the original low-dimensional target-selective spaces.

**Space Representations for Database Screening.** Next we mapped ~160,000 MDDR compounds to the selectivity set value ranges of the descriptors forming target-selective reference spaces. The results are reported in Table 4. Only ~1000 to 7000 MDDR compounds, depending on the selectivity set, matched the descriptor value ranges of the target-selective spaces (on average, ~3000 compounds). Thus, although these space representations were low-dimensional and derived only on the basis of a total of 267 GPCR antagonists, they were already highly selective, accepting on average only less than 2% of the MDDR (that contains many GPCR ligands).

We then extended the dimensionality of the target-selective spaces in order to eliminate increasing numbers of MDDR compounds from them and determine whether new target-selective compounds could be identified. Dimension extension using CA-DynaMAD was continued until only small database selection sets remained. It should be noted that dimension extension does not modify the original selectivity of these spaces. All target-selective compounds remain selected (because they represent the reference set for continued dimension extension) and all other GPCR antagonists used for space design remain deselected. However, each additional dimension extension step eliminates a subset of database compounds. The results for the last dimension extension steps are reported in Table 5. For six selectivity sets (except D4), database selection sets contained 50 or fewer MDDR compounds. For D4, 113 compounds remained because no discriminatory descriptors were available to further reduce the number of database compounds. In order to deselect nearly all MDDR compounds, the dimensionality of the original reference spaces needed to be in part significantly increased. As reported in Table 5, addition of between five (D1) and 51 (D4) descriptors was required. This was the case because random database compounds typically have much broader value ranges than selectivity sets, which generally reduces the discriminatory power of individual descriptors. Consequently, more descriptors are required.

**Table 6.** Selective Antagonists Found in the MDDR[a]

| class | correctly identified antagonists | confirmed selective antagonists |
|---|---|---|
| 5HT1a | 22 (49) | **1** [5HT1a over D1, D2, and 5HT2] |
| | | **2** [5HT1a over Alpha1] |
| 5HT2a | 6 (50) | **1** [5HT2a over Alpha1 and D2] |
| | | **1** [5HT2a over 5HT1a, D4, and Alpha1] |
| Alpha1 | 9 (45) | **1** [Alpha1 over D2 and 5HT2] |
| D1 | 11 (47) | **1** [D1 "over other GPCRs"] |
| | | **1** [D1 over D2] |
| D2 | 4 (44) | **1** [D2 over D3 and D4] |
| | | **1** (for 5HT1a over D2 and Alpha1) |
| D3 | 5 (48) | |
| D4 | 44 (113) | **3** [D4 over D2] |
| | | **1** [D4 over 5HT1a and D2] |

[a] In the MDDR entries, "activity" fields were inspected to determine the number of correctly identified antagonists among the total number of selected MDDR compounds (in parentheses) and "action" fields to obtain selectivity information (confirmed selective antagonists). The numbers of confirmed selective antagonists with different selectivity profiles are given in bold, and the selectivity profiles are reported in brackets. The profile of the only antagonist with incorrect selectivity that was detected (for D2) is reported in parentheses. For the majority of compounds, selectivity information was not available. For example, the row for "5HT1a" needs to be interpreted as follows: 22 of 49 selected MDDR compounds were designated 5HT1a antagonists ("activity"), and for three of those selectivity information was provided ("action").

However, through adaptive descriptor scoring, highly selective chemical references spaces were obtained for each of the seven selectivity sets that eliminated almost all MDDR compounds.

**Identification of Selective GPCR Antagonists.** We then analyzed the small MDDR selection sets in order to determine whether the CA-DynaMAD calculations identified selective GPCR antagonists. Therefore, the biological annotation fields of all selected MDDR compounds were studied. The results are summarized in Table 6. For each selectivity set, a number of antagonists (between four and 44) with confirmed activity against the selection set target were correctly detected but only for a total of 14 of these compounds selectivity information was available. However, 13 of these 14 compounds belonging to six selectivity sets were found to have correct selectivity profiles. Only a single false-positive compound was found for D2 (with selectivity for 5HT1a over D2 and Alpha1). Thus, reference spaces generated with CA-DynaMAD were successfully applied to identify target-selective antagonists in the MDDR, suggesting that these space representations might have considerable potential for practical applications.

**Concluding Remarks.** The design of appropriate chemical reference spaces is a key requirement for many applications in chemoinformatics and computer-aided medicinal chemistry. DynaMAD was originally developed to design compound class-directed chemical space representations and utilize them in virtual screening. With the introduction of CA-DynaMAD, we have further improved the descriptor scoring and dimension extension functions, leading to an increase in compound recall while reducing the number of descriptor variables. Going beyond virtual screening, CA-DynaMAD was successfully applied to design target-selective reference spaces for antagonists of seven biogenic amine GPCRs. The computational analysis and prediction of target selectivity is more complicated than distinguishing active from inactive compounds. In our study, highly resolved target-selective spaces that effectively deselected database compounds were successfully derived and used to identify other target-selective GPCR antagonists. Therefore, the generation of target-selective reference spaces using CA-DynaMAD complements and extends currently available approaches to chemical space design. The utility of CA-DynaMAD to identify individual descriptors that discriminate between antagonists with different selectivity has also been demonstrated. Such descriptors can be used in many other computational applications, for example, QSAR analysis. Taken together, our findings suggest that the design of "selectivity spaces" should merit further investigation and have significant potential for practical applications. Future studies will aim at generating target-selective reference spaces for additional target classes and applying them in the search for ligands with differential selectivity profiles.

## REFERENCES AND NOTES


(1) *Concepts and Applications of Molecular Similarity*; Johnson, M., Maggiora, G. M., Eds.; John Wiley & Sons: New York, 1990.
(2) Stahura, F. L.; Bajorath, J. Partitioning methods for the identification of active molecules. *Curr. Med. Chem.* **2003**, *8*, 707–715.
(3) Stahura, F. L.; Bajorath, J. New methodologies for ligand-based virtual screening. *Curr. Pharm. Des.* **2005**, *11*, 1189–1202.
(4) Martin, Y. C. Diverse viewpoints on computational aspects of molecular diversity. *J. Comb. Chem.* **2001**, *3*, 231–250.
(5) Schnur, D.; Beno, B. R.; Good, A.; Tebben, A. Approaches to target class combinatorial library design. *Methods Mol. Biol.* **2004**, *275*, 355–378.
(6) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors, in Methods and Principles in Medicinal Chemistry - Volume 11*; Mannhold, R., Kubinyi, H., Timmerman, H., Eds.; Wiley: New York, 2000.
(7) Livingstone, D. J. The characterization of chemical structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195–209.
(8) Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882–894.
(9) Lipinski, C.; Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **2004**, *432*, 855–861.
(10) Pearlman, R. S.; Smith, K. M. Metric validation and the receptor-relevant subspace concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28–35.
(11) Agrafiotis, D. K.; Lobanov, V. S.; Salemme, R. F. Combinatorial informatics in the post-genomics era. *Nat. Drug Discovery Rev.* **2002**, *1*, 337–346.
(12) Xue, L.; Bajorath, J. Accurate partitioning of compounds belonging to diverse activity classes. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 757–764.
(13) Godden, J. W.; Bajorath, J. Chemical descriptors with distinct levels of information content and varying sensitivity to differences between selected compound databases identified by SE-DSE analysis. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 87–93.
(14) Pearlman, R. S.; Smith, K. M. Novel software tools for chemical diversity. *Perspect. Drug Discovery Des.* **1998**, *9*, 339–353.
(15) Joachims, T. Making large-scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning*; Schölkopf, B., Burges, C., Smola, A., Eds.; MIT Press: Cambridge, MA, 1999.
(16) Warmuth, M. K.; Liao, J.; Rätsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. Active learning with support vector machines in the drug discovery process. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 667–673.
(17) Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1882–1889.
(18) Jorissen, R. N.; Gilson, M. K. Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.* **2005**, *45*, 549–561.
(19) Godden, J. W.; Furr, J. R.; Xue, L.; Stahura, F. L.; Bajorath, J. Molecular similarity analysis and virtual screening in binary-transformed chemical descriptor spaces with variable dimensionality. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 21–29.

(20) Eckert, H.; Bajorath, J. Determination and mapping of activity-specific descriptor value ranges (MAD) for the identification of active compounds. *J. Med. Chem.* **2006**, *49*, 2284–2293.
(21) Eckert, H.; Vogt, I.; Bajorath, J. Mapping algorithms for molecular similarity analysis and ligand-based virtual screening: design of DynaMAD and comparison with MAD and DMC. *J. Chem. Inf. Model.* **2006**, *46*, 1623–1634.
(22) Vogt, I.; Ahmed, H. E. A.; Auer, J.; Bajorath, J. Exploring structure-selectivity relationships of biogenic amine GPCR antagonists using similarity searching and dynamic compound mapping *Mol. Diversity* 2008, in press; available in the online section of the journal.
(23) Bajorath, J. Analysis of ligand relationships within target families *Curr. Opin. Chem. Biol.* 2008, in press; available in the online section of the journal.
(24) Molecular Operating Environment (MOE); Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2007.
(25) Irwin, J. J.; Shoichet, B. K. ZINC -a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
(26) MDL Drug Data Report (MDDR; version 2005.2, Oct 2005); Symyx Software: San Ramon, CA, 2005.
(27) Labute, P. Derivation and applications of molecular descriptors based on approximate surface area. *Methods Mol. Biol.* **2004**, *275*, 261–278.
(28) Balaban, A. T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **1982**, *89*, 399–404.
(29) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity - a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219–3228.
(30) Wildman, S.; Crippen, G. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868–873.
(31) Hall, L. H.; Kier, L. B. The molecular connectivity chi indexes and kappa shape indexes in structure-property modeling. *Rev. Comput. Chem.* **1991**, *2*, 367–422.

CI800106E