Mapping Algorithms for Molecular Similarity Analysis and Ligand-Based Virtual Screening: Design of DynaMAD and Comparison with MAD and DMC

Hanna Eckert,† Ingo Vogt,† and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113 Bonn, Germany

Received March 13, 2006

Here, we introduce the DynaMAD algorithm that is designed to map database compounds to combinations of activity-class-dependent descriptor value ranges in order to identify novel active molecules. The method combines and extends key features of two previously developed algorithms, MAD and DMC. These methods were first described as compound-mapping algorithms for large-scale virtual screening applications. DynaMAD and DMC operate in chemical spaces of stepwise increasing dimensionality. However, in contrast to DMC, which utilizes binary transformed descriptors, DynaMAD uses unmodified descriptor value distributions. The performance of these mapping methods was compared in detail in virtual screening trials on 24 different compound activity classes against a background of about 2 million database compounds. In these calculations, all three approaches produced results of considerable predictive value, and the enrichment of active molecules in small selection sets consisting of only about 20 or fewer database compounds emerged as a common feature. Furthermore, mapping methods were capable of recognizing remote molecular similarity relationships. Overall, DynaMAD performed better than MAD and DMC, producing average hit and recovery rates of 55% and 33%, respectively, over all 24 classes. Taken together, our findings suggest that dynamic compound mapping to combinations of activity-class-selective descriptor settings has significant potential for molecular similarity analysis and ligand-based virtual screening.

1. INTRODUCTION

Virtual screening methods can be divided into target structure-dependent[1−3] and ligand-based[4−6] approaches. Ligand-based virtual screening (LBVS) uses small molecules with known activity as templates for database analysis and the identification of novel hits by application of various molecular similarity[6,7] and structure−activity relationship methods.[8,9] Molecular-similarity-based LBVS methods can generally be divided into two categories, similarity searching and compound classification approaches.[4] Computational tools for similarity searching[10] include, for example, diverse molecular fingerprints,[11−13] substructures,[14] and suitable molecular graph[15] or shape[16] representations. Compound classification methods with relevance for LBVS include classical clustering methods[17,18] and further advanced cluster algorithms;[19,20] statistical[21,22] and dimension reduction-based[23,24] partitioning methods;[25] and machine learning approaches such as neural networks,[26] support vector machines,[27] and other kernel-based methods.[28]

LBVS methods generally depend on the use of molecular structure and property descriptors[29−31] for the generation of chemical reference spaces,[31] and chemical space design is generally recognized as a crucial step for the successful application of these methods.[31,32] An important factor in the design is the dimensionality of chemical space, and the generation of low-dimensional chemical reference spaces has become a paradigm for molecular similarity analysis, compound partitioning, or library design[23,31,32] (although there are exceptions such as support vector machines[27] that project compounds into high-dimensional space representations). Another critical factor is descriptor selection. Because literally thousands of different chemical descriptors are available,[29] it is generally difficult to estimate or determine which descriptors might perform best in a specific application.[4,31] Multidimensional scaling techniques,[32] machine learning methods such as genetic algorithms,[33] and also information theory[34] have been employed to aid in descriptor selection, but the systematic identification of molecular descriptors that selectively respond to specific compound activities continues to be an area of high interest in the chemoinformatics field, although progress has been fairly limited in recent years.[4,6]

A number of LBVS methods do not scale well with the rapidly increasing size of compound libraries, for example, those that depend on the evaluation and comparison of molecular conformations such as pharmacophore-based[8] or multidimensional QSAR[35] approaches or methods that depend on pairwise compound or distance comparisons in chemical space such as many clustering algorithms.[17,25] At present, virtually formatted screening libraries typically contain millions of molecules, and computational efficiency thus becomes another important aspect for the development of LBVS methods.

Efforts in our laboratory have focused on the development of what we call mapping algorithms for molecular similarity analysis and virtual screening, which add to the spectrum of currently available LBVS methods. The first mapping algorithm we introduced was termed *D*ynamic *M*apping of *C*onsensus positions (DMC).[36,37] Recently, we developed another mapping method called *M*apping to *A*ctivity-class-specific *D*escriptor value ranges (MAD).[38] Herein, we report the design and evaluation of the *Dyna*mic MAD (DynaMAD)

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

† The contributions of these authors should be considered equal.

10.1021/ci060083o CCC: $33.50
Published on Web 06/01/2006

algorithm, which combines the dimension extension approach of DMC with the activity-class-specific descriptor representation and selection of MAD and incorporates more refined descriptor value range determination and compound mapping functions. Furthermore, the performance of these methods is compared in detail in simulated virtual screening calculations targeting 24 different compound activity classes added to a publicly available screening database containing ~2 million compounds. In the following sections, we describe unique features of these algorithms, provide methodological details, introduce DynaMAD, and present the results of comparative virtual screening trials.

2. CHARACTERISTIC FEATURES OF MAPPING ALGORITHMS

2.1. Design Principles. These methods are designed to address difficulties involved in descriptor selection and chemical space design and to facilitate the screening of large compound databases. Regardless of the specifics of each algorithm, sets of active compounds are used to select descriptors from large pools that best define unique and activity-dependent positions in chemical space. In addition, mapping algorithms deliberately depart from the paradigm of low dimensionality, as they operate in high-dimensional reference spaces or at least spaces of progressively increasing dimensionality. Molecular property descriptors are employed either in binary or in range-keyed form. Activity-dependent consensus positions in chemical space or combinations of activity-specific descriptor value ranges are used to map database compounds and identify those molecules that are most similar to known actives and thus represent potential hits.

2.2. DMC and Simplified Descriptor Spaces. DMC is designed to identify consensus positions for classes of active compounds in descriptor spaces of stepwise increasing dimensionality and thereby distinguish active molecules from other database compounds. These chemical reference spaces are generated using property descriptors that have been simplified through binary descriptor transformation, which converts descriptors having discrete or continuous value ranges into a binary format.[39] This is achieved by calculating the statistical medians of the value distributions of the descriptors in the screening database. Each compound is assigned a "1" for a descriptor if its value is larger than the median or a "0" if it is equal to or smaller than the median. A consensus position is defined by a descriptor vector in chemical space composed of those bit settings that are identical for all compounds belonging to an activity class. Dimension extension is achieved by calculating consensus positions that no longer require identical descriptor settings for all known actives but permit a certain percentage of descriptor (bit) variability, for example, 10%, 20%, and 30%. Permitting descriptor variability effectively increases the number of binary descriptors constituting a consensus position and, thus, the resolution of the space representation. During dimension extension, the number of database compounds that map to activity-dependent consensus positions usually sharply decreases and only similar compounds sharing the signature bit strings are retained.[36] These database compounds are considered potential hits.

Although binary transformed descriptors are per se low-resolution formulations, DMC calculations have shown that combinations of transformed descriptors become highly selective database search tools. Even for large compound databases (e.g., with more than a million molecules), DMC eliminated most of the database molecules from activity-dependent consensus positions after only the second or third dimension extension step, and in pilot calculations on five therapeutically relevant activity classes, DMC produced hit rates of up to ~70%.[36] The DMC methodology has also been extended by adding a potency scaling function to tune database search calculations toward the recognition of increasingly potent hits.[37] On the basis of its design, to facilitate compound selection and deselection during dimension extension, DMC combines elements from both bit string and partitioning methods.[36]

2.3. MAD and Activity-Dependent Descriptor Value Ranges. The development of MAD was catalyzed by our findings that many descriptors adopt narrow value ranges for sets of active compounds that differ from their value distribution in screening databases.[38] Given these results, we designed an algorithm to map database molecules to multiple activity-specific descriptor value ranges and identify potential hits by focusing on those compounds that closely match these ranges. Thus, similar to DMC, MAD is also a mapping technique in chemical reference space. However, in contrast to DMC, it is "static" because its dimensionality is given and not modified. Moreover, MAD utilizes original descriptor values and not binary transformed or otherwise simplified descriptor representations. MAD systematically compares descriptor value ranges of compound activity classes (e.g., known inhibitors, agonists, or antagonists) with those of screening database compounds in order to identify descriptors having activity-class-specific settings. By contrast, DMC calculations capture such effects implicitly by generating descriptor bit strings of increasing length that become a signature of an activity class. The MAD formalism requires activity-class-specific or selective descriptors to display value distributions for active compounds that detectably depart from the database distribution. This is the case if active molecules adopt values rarely seen in database compounds such as value ranges at the upper or lower end of the range for these compounds or, alternatively, if they produce appropriately positioned narrow value ranges within the database range. MAD utilizes a scoring function to quantify those effects and rank descriptors according to decreasing sensitivity for active compounds. This makes it possible to select a prespecified number of descriptors and use them in combination for the mapping of compounds to activity-class-sensitive descriptor value ranges.[38]

3. METHODOLOGICAL DETAILS

3.1. DMC Descriptor Scoring. Once binary transformed descriptors are calculated and each active compound is assigned a descriptor bit string, a descriptor scoring function is applied using the mean (bitMean) of the class bits as an indicator of bit variability:

$$\text{score} = |0.5 - \text{bitMean}|$$

On the basis of this simple function, top scores of 0.5 are achieved if all bits are set identically or a minimal score of 0 is achieved if half of the bits are set to 0 and the other half to 1. Thus, descriptors with low bit variability have high

scores, and their values fall mostly on the same side of the database median, which potentially indicates a class-specific value setting and feature. After the calculation of scores, descriptors are assigned to different layers. Layer 0 defines the initial consensus position and starts with the top descriptors having the score 0.5 (no bit variability). All other layers are built, for example, by taking "0.1" steps; that is, descriptors with scores ≥ 0.4 are assigned to layer 1, those with scores ≥ 0.3 to layer 2, and so on. Only descriptors with the minimal score 0 (bitMean = 0.5) are omitted, because they show no class-specific tendency. Thus, dimension extension in DMC, as used here, is facilitated by allowing 10% additional bit variability for a given compound set per extension step. Database compounds are mapped to consensus positions resulting from dimension extension through different descriptor layers by calculating corresponding descriptor bit strings. If they match the consensus position (or string), they are retained; otherwise, they are discarded.

3.2. MAD Descriptor Ranking and Compound Mapping. Briefly, for each descriptor, one calculates the 25% quantile ($q^{0.25}$) and the 75% quantile ($q^{0.75}$) of its database value distribution and distinguishes the central 50% of the value distribution (termed centralRange) from the highest 25% and lowest 25% of the database values. For compounds of an activity class, the minimum (classMin) and maximum (classMax) value of each descriptor is determined and the descriptor value range (exactRange) of the activity class is defined as exactRange = [classMin, classMax]; the size of the value range is sizeRange = classMax − classMin.

On the basis of this formalism, it is then possible to score descriptors. For example, one possibility is that the value range of an activity class falls within the centralRange or overlaps with it: classMax $\geq q^{0.25}$ and classMin $\leq q^{0.75}$. Then, the descriptor score is calculated as

$$\text{score} = \frac{q^{0.75} - q^{0.25}}{\text{sizeRange}}$$

Other possibilities are that the value range of an activity class falls completely below the low-25% quantile or above the high-25% quantile of the database distribution, and these cases are treated with variants of the above scoring function, as described.[38] When the scoring function is applied, descriptors obtain scores smaller than 1 if the database value range between the 25% and 75% quantiles is smaller than the class value range. This corresponds to a situation where usually more than 50% of the database molecules match the class value range of the descriptor, and this descriptor would not be regarded as an activity-sensitive one and would not be used for mapping. Descriptor scores greater than 1 are obtained when less than half of the database molecules match the value range of an activity class. The fewer database molecules that match, the higher the score and the likelihood of activity-class-specific descriptors increases. Thus, on the basis of their scores, descriptors are ranked according to potential class specificity.

To identify new active compounds by virtual screening, database molecules are mapped to unmodified or slightly extended descriptor value ranges of known active molecules. Database compounds matching value ranges of selected descriptors are ranked using a simple similarity function that divides the number of matched value ranges by their total number. Value range extension is applied in order to increase the ability to recognize compounds with similar activity but increasingly diverse structures, a process often called "lead hopping",[40] which represents an important goal of virtual screening analysis.[4] In the original MAD implementation, lead-hopping potential was further increased by the introduction of an expansion function depending only on the class range size and the number of bait molecules. For these purposes, we define

$$\text{dExp} = \frac{\text{sizeRange}}{|\text{Baits}| - 1}$$

"Baits" represents the set of active compounds used as templates for virtual screening. These compounds are also used to determine exactRange. This leads to the following expanded range for compound mapping:

$$\text{expandedRange} = [\text{classMin} - \text{dExp}, \text{classMax} + \text{dExp}]$$

In benchmark calculations, MAD produced promising results. When utilized to mine a database of bioactive molecules for six specific activities, hit rates between ~15% and 79% were obtained.[38]

3.3. Specifics of DynaMAD. In DynaMAD, the descriptor scoring function and value extension mechanisms are modified by using mapping probabilities of class value ranges instead of value range sizes. By doing so, descriptors with arbitrary and, especially, abnormal value distributions can be treated uniformly in a predictable manner. The second major modification in DynaMAD is the introduction of dynamic descriptor selection instead of a single-step or static selection routine. In MAD, descriptor selection was carried out during the initial step, leading to a constant descriptor number for mapping. By contrast, in DynaMAD, the number of selected descriptors is increased in a stepwise manner and alternates with compound mapping, which is reminiscent of determining descriptor consensus positions during the dimension extension steps of DMC. Thus, during DynaMAD calculations, descriptor sets can be augmented individually for each activity class until optimal descriptor numbers are obtained.

3.3.1. Value Distribution of Descriptors. For the compound database used as the source for virtual screening, two different distribution scenarios are distinguished for descriptors, and their value distributions are represented as follows:

(a) For descriptors having a discrete value distribution with up to ~100 different values in the database, the relative frequency of each descriptor value is calculated. At the lower and upper ends of the value distribution, where only small relative frequencies are obtained, descriptor values are pooled and their relative frequencies are summed up to maximal 1% overall relative frequency. Table 1 reports an example illustrating how the value distribution of the descriptor *a_nC* (number of carbon atoms) is represented accordingly.

(b) For descriptors with a continuous value range or a discrete value distribution containing significantly more than 100 or so values (pseudo-continuous distribution), we divide the value distribution into "mini ranges" with an overall relative frequency of occurrence of maximal 1% of the values obtained for database compounds. Single descriptor values with a relative frequency greater than 1% are stored individually. Mini ranges can be easily calculated by

Table 1. Example of a Discrete Value Distribution[a]

descriptor value	relative frequency	descriptor value	relative frequency
≤7	0.008	18	0.098
8	0.006	19	0.094
9	0.010	20	0.087
10	0.016	21	0.075
11	0.022	22	0.062
12	0.030	23	0.049
13	0.040	24	0.038
14	0.054	25	0.028
15	0.068	26	0.019
16	0.083	27	0.011
17	0.093	≥28	0.009

[a] Value distribution for descriptor *a_nC* (number of carbon atoms in a molecule). The relative frequencies of the smallest descriptor values 0−7 and the descriptor values greater than 28 were summed up to obtain an overall relative frequency of maximal 1%.

Table 2. Example for a Continuous Value Distribution Captured in Mini Ranges[a]

mini range	relative frequency	mini range	relative frequency
[0.000, 0.000]	0.333	(17.048, 17.743)	0.000
(0.000, 5.683)	0.000	[17.743, 17.743]	0.032
[5.683, 5.683]	0.369	(17.743, 18.842)	0.000
(5.683, 9.421)	0.007	[18.842, 18.842]	0.012
[9.421, 9.421]	0.053	(18.842, 23.425)	0.006
(9.421, 11.365)	0.002	[23.425, 23.425]	0.019
[11.365, 11.365]	0.099	(23.425, 26.950)	0.009
(11.365, 15.104)	0.005	[26.950, 28.865)	0.008
[15.104, 15.104]	0.020	[28.865, 39.620)	0.010
(15.104, 17.048)	0.001	[39.620, 73.169]	0.002
[17.048, 17.048]	0.013		

[a] Value distribution for descriptor *vsa_don*, which approximates the sum of van der Waals surface areas of hydrogen-bond donors. Single descriptor values (like 5.683 or 23.425) having a relative frequency greater than 1% are stored individually. Otherwise, descriptor values are combined to mini ranges having a total relative frequency of maximal 1%.

combining adjacent descriptor values until an overall relative frequency of 1% is obtained. Table 2 gives an example for the approximation of continuous value distributions by the introduction of mini ranges.

3.3.2. Descriptor Scoring. The scoring function enables the comparison of descriptors and is based on the assessment of specificity of descriptor value ranges for an activity class. The descriptor class value range between the minimum (classMin) and maximum (classMax) values of a class is thought to be increasingly specific as fewer database compounds are matched. Therefore, the DynaMAD scoring function is designed to directly relate descriptor scores to mapping probabilities of class ranges:

$$\text{score} = [1 - P(\text{classMin} \le X \le \text{classMax})]100$$

By applying this function, a maximum descriptor score of 100 can be achieved (i.e., no database compound matches the class range) and a minimum score of 0 (i.e., all compounds match the class range). Mapping probabilities required for score calculation are obtained by the addition of (a) relative frequencies of discrete values x_i for descriptors with discrete distributions

$$P(\text{classMin} \le X \le \text{classMax}) = \sum_{\text{classMin} \le x_i \le \text{classMax}} P(X = x_i)$$

and (b) relative frequencies of mini ranges $[x_{i1}, x_{i2}]$

$$P(\text{classMin} \le X \le \text{classMax}) = \sum_{\text{classMin} \le x_{i2}; \text{classMax} \le x_{i1}} P(X \in [x_{i1}, x_{i2}])$$

Mini ranges either fall within class value ranges or overlap with them. The maximum possible absolute error due to the introduction of this approximation for overall mapping probabilities is 0.02, but the error is usually negligible or zero for discrete value distributions.

3.3.3. Descriptor Value Ranges for Compound Mapping. Descriptor value ranges of activity classes are used for the mapping of database compounds. To favor the recognition of molecules with similar activity but increasingly diverse structures and, thus, the probability of recognizing remote similarity relationships (lead hopping), class value ranges are extended in an analogous manner to MAD. For this purpose, two different expansion functions can be applied in DynaMAD. Both functions make the extent of value range expansion Δp dependent on the mapping probability p. The consequences of using bait sets of varying size are implicitly accounted for by resulting changes in mapping probabilities.

$$\text{Function 1: } \Delta p = \frac{1 - p}{100p + 10}$$

$$\text{Function 2: } \Delta p = \frac{1 - p}{8}$$

These two alternative functions differ in the degree of expansion. The constant 10 in function 1 ensures that the minimum mapping probability following expansion is 10%. Because function 2 should principally lead to a larger expansion than function 1, a constant of 8 is used, which corresponds to a minimum mapping probability of 12.5%. For increasing mapping probabilities, Δp decreases. Function 1 only leads to a significant expansion in the presence of low mapping probabilities (essentially causing no expansion at mapping probabilities above 0.5), whereas function 2 constantly reduces the magnitude of expansion with increasing mapping probabilities. Thus, expansion function 1 is more akin to a correction function for low mapping probabilities than function 2. Given Δp, class value ranges are augmented on the lower and upper ends by including the next discrete value or mini range until the total additional mapping probability Δp is included. For symmetrical expansion, the enlarged class value range is calculated as

$$\left[\text{classMin} - \frac{\Delta p}{2}, \text{classMax} + \frac{\Delta p}{2}\right]$$

Following value range expansion, scores are updated to be consistent with the new mapping probabilities:

$$\text{updatedScore} = \text{score} - \Delta p 100$$

3.3.4. Descriptor Classification. On the basis of the obtained DynaMAD scores, descriptors are assigned to different categories or layers. For this purpose, the available descriptor score range (0 − 100) is divided into equally sized subranges, and all descriptors falling within the same subrange form one layer. As an example, score subranges

of size 5, as used here, produce a total of 20 layers, each representing a dimension extension step for building DynaMAD descriptor spaces. The layer representing descriptors with the highest scores (≥95) is given number 0, the next (≥90) number 1, and so on. Descriptor layer 0 in DynaMAD corresponds to the initially defined consensus position in DMC.

3.3.5. Dynamic Compound Mapping. The mapping process starts with layer 0 and then proceeds to dimension extension. Database compounds are regarded as similar to baits if their descriptor values fall into the class value ranges belonging to the current layer. Only compounds that match all of the value ranges qualify for mapping to the next layer; others are discarded. In this next round, remaining compounds also have to match the class value ranges of the associated descriptors in order to qualify for further mapping. Thereby, the number of class value ranges for mapping is increased in a stepwise manner, and fewer database compounds map to the baits until only a small number of similar database molecules remain. Thus, the compound mapping strategy of DynaMAD is distinct from that of MAD but parallels the dimension extension steps in DMC, albeit utilizing completely different descriptor formulations and selection methods.

3.3.6. Parameter Optimization. For individual compound activity classes, parameter settings for DynaMAD calculations are optimized with respect to value range expansion and statistical outliers of bait set descriptor values. Calculations are carried out either without value range expansion or by applying expansion functions 1 or 2 and, in addition, either with unmodified bait descriptor value sets or after removing the minimum and maximum descriptor values from each set. The removal of these values is of potential relevance for the treatment of highly heterogeneous bait compound sets where single compounds might produce "outlier" values. DynaMAD virtual screening calculations for each class are carried out using optimized parameter settings and MAD reference calculations applying the previously established value range expansion protocol,[38] as outlined above, without further modifications. For practical virtual screening applications, DynaMAD would be trained on a given compound activity class as described above in order to identify preferred parameter settings.

4. MATERIALS AND CALCULATIONS

4.1. Activity Classes, Descriptors, and Database Compounds. As a source database for our studies, we use the publicly available compound collection ZINC containing ~2.01 million molecules from various vendor sources.[41] In our analysis, all ZINC compounds are considered inactive (and thus potential false positives), although it is anticipated that ZINC contains hits for at least some of the activity classes studied here. In our calculations, 155 1D, 2D, and implicit 3D descriptors implemented in the Molecular Operating Environment (MOE)[42] are used as a basis set or pool. In principle, as many descriptors as possible can be included in the basis set, because mapping algorithms select preferred descriptors. Our only requirement for descriptor preselections is that they should not depend on hypothetical compound conformations. Descriptor values for activity classes and database compounds are calculated with MOE.

Table 3. Activity Classes for Virtual Screening Trials

Activity Classes Assembled from the Literature		
class designation	biological activity	number of compounds
5HT	serotonin receptor ligands	21
ACE	acetylcholine esterase inhibitors	17
BEN	benzodiazepine receptor ligands	22
CAE	carbonic anhydrase II inhibitors	22
COX	cyclooxygenase-2 (Cox-2) inhibitors	17
H3E	histamine H3 antagonists	21
HIV	HIV protease inhibitors	18
TKE	tyrosine kinase inhibitors	20

Activity Classes Assembled from the MDDR		
class designation	biological activity	number of compounds
ANA	angiotensin II antagonists	45
CHO	cholesterol esterase inhibitors	30
DD1	dopamine D1 agonists	30
DIR	dihydrofolate reductase inhibitors	30
EDN	endothelin antagonists	32
ESU	estrone sulfatase inhibitors	35
GLY	glycoprotein IIb/IIIa receptor antagonists	34
INO	inosine monophosphate dehydrogenase inhibitors	35
KAP	κ agonists	25
LAC	β-lactamase inhibitors	29
LDL	LDL receptor upregulators	30
MEL	melatonin agonists	25
SQS	squalene synthetase inhibitors	42
THI	thiol protease inhibitors	34
THR	thromboxane antagonists	33
XAN	xanthine oxidase inhibitors	35

For DMC analysis, descriptors are binary-transformed using the medians of their ZINC value distributions.

Comparison of DynaMAD with MAD and DMC is carried out on 24 different activity classes containing between 17 and 45 compounds, as reported in Table 3. Eight of these classes were originally assembled from the literature for a partitioning analysis,[43] and 16 additional classes were extracted from the Molecular Drug Data Report (MDDR)[44] for our present analysis. To assemble these new activity classes, MDDR compounds sharing the same activity were filtered to ensure compliance with lead- or druglike standards[45] with respect to the molecular weight and logP(o/w) (logarithm of the octanol/water partition coefficient) and subsequently clustered using the publicly available set of 166 MACCS structural fragment descriptors[46] and a hierarchical clustering routine implemented in MOE. Molecules from nonsingleton clusters were randomly selected for our activity classes. This was done in order to avoid the inclusion of analogue series in these classes and to ensure structural diversity within classes.

4.2. Virtual Screening Trials and Performance Measures. For each of the 24 activity classes, 100 sets of 10 compounds each are randomly selected. Each set is taken once as a bait set for the calculation of descriptor scores, assignment of descriptors to different layers or dimensions, and determination of specific value ranges (MAD and DynaMAD) or consensus bit patterns (DMC). The remaining active compounds (between 7 and 35) are added to ZINC as potential hits. Thus, for each activity class, 100 different search calculations are carried out, thereby limiting bias due to chance effects in the selection of baits and potential hits. As performance measures for MAD, DynaMAD, and DMC, hit rates (HR; number of selected active compounds relative to selected database molecules) and recovery rates (RR; number of selected active compounds relative to the total

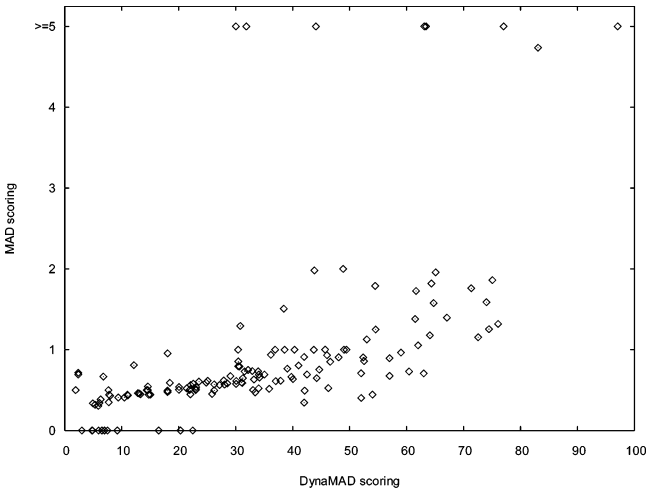

Figure 1. MAD versus DynaMAD descriptor scoring. As an example, descriptor scores obtained with MAD and DynaMAD scoring functions are compared for activity class MEL. DynaMAD scores range from 0 to 100, whereas MAD scores have no fixed upper limit. MAD scores equal to or greater than 5 are considered a high scoring range. Thus, the upper-right corner of the graph contains descriptors with consistently high scores, whereas the lower-left corner contains those that score low with both functions (and would not be selected).

number of potential database hits) are calculated for each dimension extension level and averaged over 100 independent trials. In addition, we define HRR as a combined hit and recovery rate, calculated as the geometric mean of HR and RR (i.e., square root of their product). Thus, HRR represents a consensus measure for the outcome of virtual screening calculations. We calculate the geometric mean rather than the arithmetic average in order to better address the situation at the beginning and end of mapping calculations. Prior to dimension extension, the number of database compounds can be relatively large, resulting in artificially high recovery rates and low hit rates. At the end of mapping calculations, compound selection sets are small, hit rates are increased, and recovery rates decreased. Therefore, geometric means provide a better rate balance at these points, in particular, at the beginning of the calculations.

5. RESULTS AND DISCUSSION

5.1. Descriptor Scoring and Calculation Parameters. The assessment of activity-class-specific descriptor settings represents a major difference between the MAD and DynaMAD methods. In contrast to MAD, the DynaMAD scoring function directly relates descriptor scores to mapping probabilities of class ranges. We find that both scoring schemes significantly differ in the evaluation of descriptor relevance. Figure 1 shows a representative example. In this case, only two to three descriptors score equally high using both functions (upper-right corner), but many others score low (consistent with the fact that the majority of descriptors do not selectively respond to compound-class-specific features). For MAD, descriptor scores greater than 1 begin to display class-specific tendencies,[38] and for DynaMAD, scores greater than 50 indicate that progressively fewer than half of the database compounds map to class ranges. In Figure 1, a number of descriptors show relative differences in MAD versus DynaMAD scoring, leading to the selection of different descriptor sets for compound mapping.

Table 4. Preferred DynaMAD Calculation Parameters[a]

	preferred parameters			preferred parameters	
activity class	removal of outliers	expansion function	activity class	removal of outliers	expansion function
5HT	−	1	H3E	−	2
ACE	+	2	HIV	−	1
ANA	−	1	INO	+	
BEN	−	2	KAP	+	1
CAE	−		LAC	−	
CHO	−	2	LDL	+	2
COX	−		MEL	+	
DD1	−	2	SQS	−	1
DIR	+		THI	−	2
EDN	−		THR	+	
ESU	+		TKE	−	
GLY	−	2	XAN	+	2

[a] For each activity class, the parameter settings that produced the best results are reported. Results varied dependent on whether expansion functions were applied and minimum and maximum descriptor values (potential "outliers") were removed or not.

Additional features of DynaMAD compared to those of MAD include the choice of alternative expansion functions and outlier value removal. Preferred combinations of these parameters were identified for each activity class and are reported in Table 4. The best parameter settings differed for activity classes, and no generally preferred parameter combination was identified. However, outlier removal was not required in the majority of test cases, and for 10 classes, value range expansion did not improve the results. For 9 of the 14 remaining classes that benefited from class range expansion for compound mapping, expansion function 2 was preferred over function 1. Taken together, these findings indicate that, if possible, DynaMAD parameter settings should be examined for individual activity classes.

5.2. DynaMAD Performance in Virtual Screening Trials and Comparison with MAD. DynaMAD test calculations on our 24 activity classes produced encouraging results reported in Table 5. In each case, only between 7 and 35 active compounds were added as potential hits to 2 million ZINC compounds (that were all considered false positives). With one exception (activity class THI), DynaMAD achieved hit rates between 24% and 100% and recovery rates between 15% and 97%. For half of the activity classes, hit rates greater than 50% were obtained with a minimum of 20−25% recovered compounds. MAD also produced overall good results with a maximum hit rate of 71% and a recovery rate of 85%. Here, eight classes displayed hit rates lower than 10% and six classes displayed hit rates higher than 50%. MAD calculations were more limited in the recovery of compounds than DynaMAD, although five classes yielded recovery rates greater than 50%. For 14 other classes, less than 10% of the compounds were recovered. It should be noted that the scoring scheme of MAD was originally developed on the basis of benchmark calculations on six sets of active compounds distinct from those studied here[38] and was not further modified for our calculations. Regardless, DynaMAD produced better hit rates than MAD on every class and better recovery rates for 21 of 24 classes, which demonstrates the potential of iterative compound mapping and further refined descriptor selection routines. Compared to MAD, DynaMAD alleviates the need to select predefined numbers of descriptors for mapping and makes it possible to focus on those descriptor contributions

Table 5. Comparison of MAD, DynaMAD, and DMC[a]

	MAD		DynaMAD				DMC			
activity class	HR [%]	RR [%]	DEL	DS	HR [%]	RR [%]	DEL	DS	HR [%]	RR [%]
5HT	15.8	14.4	7	23	52.7	26.5	0	40	0.6	48.6
ACE	59.3	84.8	1	25	98.0	51.3	0	91	91.6	52.1
ANA	60.9	17.4	5	44	62.7	21.0	1	75	25.2	40.2
BEN	12.0	10.0	10	53	44.3	29.4	0	58	3.5	14.3
CAE	9.6	8.0	5	26	56.0	27.5	0	67	1.9	45.8
CHO	7.1	3.6	18	137	44.6	15.4	0	45	8.6	15.8
COX	37.7	53.8	0	3	100.0	97.3	0	54	33.2	51.7
DD1	41.4	20.7	13	94	88.2	19.1	1	81	64.2	14.3
DIR	16.7	8.3	1	5	43.7	36.9	0	45	2.9	12.9
EDN	6.5	3.0	7	30	28.2	19.4	1	49	10.5	34.2
ESU	16.8	6.7	0	2	52.2	63.6	0	38	44.1	11.4
GLY	60.2	25.1	8	59	69.6	24.7	0	79	78.9	39.5
H3E	65.4	59.4	8	70	97.9	37.4	0	77	79.2	46.1
HIV	65.5	81.8	1	39	100.0	44.4	1	108	68.0	23.0
INO	14.8	5.9	3	8	33.4	19.2	0	40	17.8	14.2
KAP	5.0	3.3	2	11	24.0	21.7	0	80	6.0	10.0
LAC	7.3	3.9	6	10	27.0	31.3	0	50	30.5	11.2
LDL	6.8	3.4	6	33	24.8	27.2	2	58	21.4	20.9
MEL	3.1	2.1	0	6	52.8	38.8	1	87	77.3	10.4
SQS	25.2	7.9	5	39	29.3	24.9	2	101	38.6	20.1
THI	2.2	0.9	12	79	4.6	10.1	0	80	27.5	7.5
THR	14.4	6.3	1	5	52.6	34.0	1	88	47.7	11.2
TKE	71.4	71.4	1	7	100.0	61.1	1	61	56.1	15.1
XAN	14.2	5.7	6	70	45.3	19.0	0	94	18.2	11.0

[a] For MAD calculations, the top 80 descriptors on the scoring list were consistently used in each case and recovery and hit rates are reported for selection sets of 10 compounds. For DynaMAD and DMC, the combinations of hit and recovery rates resulting in the largest geometric mean (HRR) and the corresponding number of descriptors are shown. DynaMAD and DMC selection sets were similar in size, ranging from 1 to 20 compounds for DMC and 5 to 20 compounds for DynaMAD. DS stands for descriptors and DEL for dimension extension level.

having the most significant class-specific potential. For example, for eight activity classes, only 10 or fewer descriptors were selected by DynaMAD and were sufficient for high prediction accuracy. This clearly demonstrates the ability of the approach to identify and map class-specific value ranges. This is further illustrated by our finding that descriptors belonging to the highest-scoring layer alone were sufficient to produce the best results in three cases (thus, no dimension extension was required) and that a single dimension extension step was sufficient in five others. In addition, both MAD and DynaMAD achieved reasonable to high hit and recovery rates for very small compound selection sets (i.e., 10 compounds for MAD and variably sized sets for DynaMAD containing up to 15 compounds). This characteristic was first observed for DMC[36] and emerges as a common feature of these mapping algorithms, indicating high specificity in compound selection (low false-positive rates).

5.3. Comparison with DMC. Given the improved performance of DynaMAD relative to "static" MAD, we next included DMC in the comparison. The results of virtual screening trials using DMC are also summarized in Table 5. Overall, the performance of DMC was intermediate. DMC produced better results than MAD, with higher hit and recovery rates in 17 and 18 cases, respectively, but was outperformed by DynaMAD, which produced higher hit rates than DMC for 19 of 24 classes and higher recovery rates for 16 classes. Figure 2a illustrates that the magnitude of combined hit and recovery rates (HRRs) obtained for DynaMAD and DMC was comparable in many cases but that overall higher HRRs were produced with DynaMAD. While the three algorithms compared here performed similarly in a number of cases, some calculations revealed significant compound-class-specific differences between these methods. Among others, examples include activity class MEL, which could be successfully analyzed with both DynaMAD and DMC but not with MAD, or activity class DIR, which could only be successfully treated with MAD and DynaMAD but not with DMC. By contrast, DMC produced acceptable results for THI, but both MAD and DynaMAD failed. Such compound-class-specific differences in classification or search performance have been well-documented for other virtual screening approaches[4] and are thought to be a common feature of many molecular-similarity-based methods.[4,47]

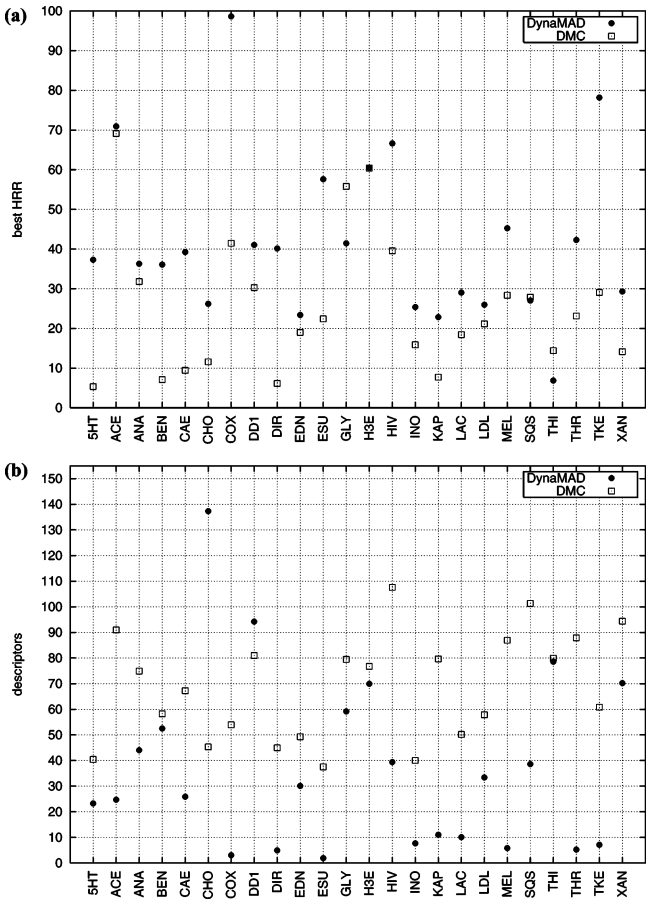

Figure 2. Virtual screening results for DynaMAD and DMC. Results obtained for each activity class are shown. In a, the best combined hit and recovery rates (HRR) are compared, and b reports the number of descriptors producing the best HRR rates.

The major differences between the DMC and DynaMAD mapping algorithms are descriptor selection and representation. DMC utilizes binary-transformed property descriptors that, as single descriptors, have much lower information content and "resolution" than descriptors with unmodified and continuous value ranges that are used by DynaMAD. However, bit strings of combinations of multiple binary-transformed descriptors become highly discriminatory in compound recognition.[36,37] This is evident from the results reported in Table 5.

5.4. Discriminating Descriptors. We have also analyzed the descriptors selected by each mapping algorithm. Figure 2b compares by activity class the number of descriptors producing the best HRRs. As would be expected, DMC calculations select on average a significantly larger number

of (simplified) descriptors than DynaMAD for the definition of activity-dependent consensus positions, but their numbers are more evenly distributed across different classes than those for DynaMAD. For 15 activity classes, initially determined "binary" consensus positions of DMC already gave the best results, and only for two classes were two subsequent dimension extension steps required (Table 5). However, Figure 2b also shows that for a number of classes DynaMAD required only 10 or fewer descriptors to produce the best HRRs, which reflects distinct class specificity.

The composition of preferred descriptor sets is comparable for DynaMAD and MAD because we adopted and further modified the descriptor selection principle of MAD for DynaMAD in order to support the identification of highly discriminatory descriptors.

When comparing the descriptor sets of DynaMAD and DMC, the following observations are made. Preferred DynaMAD descriptors usually also contribute to consensus positions in DMC. The explanation for this phenomenon is simple. High descriptor scores in DynaMAD are reached for classes with small descriptor value ranges. These small ranges fall with high probability completely below or above the database median and, thus, participate in a consensus position. Among others, we find that connectivity and shape indices and also partial charge descriptors lead to the overall best descriptor scores and often contribute to consensus positions. However, DMC selects descriptors for consensus positions irrespective of how many database compounds adopt similar descriptor values, whereas DynaMAD identifies descriptors for which the class descriptor value range is mapped by as few database compounds as possible. Thus, simple descriptors such as atom and bond counts having not much information content in the background database often participate in the definition of DMC consensus positions but are not selected by DynaMAD. As an example, for the 24 activity classes studied here, the descriptors for the number of bromine, fluorine, or chlorine atoms as well as the descriptors for the number of triple bonds and the sum of formal charges occur in DMC consensus positions for almost every class but were not selected for preferred extension levels by DynaMAD. These descriptors adopt the value zero for more than 75% of the background database and, thus, produce low descriptor scores in DynaMAD calculations.

5.5. Dimension Extension Characteristics. To compare the dynamic mapping approaches of DynaMAD and DMC in more detail, we have analyzed single search calculations and graphically compared DynaMAD and DMC hit and recovery rates during dimension extension. Table 6 reports exemplary single virtual screening runs on two activity classes. For 5HT, DynaMAD selects the first two descriptors during the third dimension extension step. Mapping database compounds to their value ranges retains 10 of 11 potential hits but already eliminates almost 2 million ZINC compounds (again demonstrating class-specific effects). The fourth dimension extension step selects six additional descriptors, retains seven potential hits, and eliminates more than 150 000 of the remaining ZINC compounds. Changes during the fifth dimension extension steps are minute, but the sixth step engages 17 descriptors and leads to another sharp reduction in compound numbers so that only a total of 23 database compounds remain, five of which are active. As discussed above, for 5HT, dimension extension is required to select crucial descriptors. By contrast, for ANA, DynaMAD identifies three descriptors prior to dimension extension, and only 1314 of more than 2 million ZINC compounds map these ranges. During the first dimension extension step, two descriptors are added and 23 active database compounds are retained together with only 97 ZINC compounds. Mapping to nine descriptors during the second extension step eliminates all ZINC compounds and retains 19 hits. Interestingly, one of these hits maps class value ranges of all of the descriptors when dimension extension is continued, which might seem surprising at first glance. However, the explanation for this is that this compound is an analogue of one of the bait molecules and, therefore, always falls within a class value range. The DynaMAD single trial results for 5HT and ANA illustrate differences in dimension extension and compound mapping profiles dependent on the activity class. Moreover, comparison with the corresponding DMC results, also reported in Table 6, reveals intrinsic differences between these dimension extension methods. Because DMC utilizes simplified descriptors, a much larger number of descriptors than that for DynaMAD define the initial activity-dependent consensus positions: 39 descriptors for 5HT and 81 for ANA. However, these descriptor combinations are highly discriminatory. In both cases, about half of the potential hits match the consensus positions and very small numbers of

Table 6. Single Trial Results for DynaMAD and DMC[a]

DynaMAD

	5HT					ANA				
DEL	DS	ADC	DC	HR [%]	RR [%]	DS	ADC	DC	HR [%]	RR [%]
0	0	11	2 066 541	0.0	100.0	3	23	1314	1.7	65.7
1	0	11	2 066 541	0.0	100.0	5	23	97	19.2	65.7
2	0	11	2 066 541	0.0	100.0	9	19	0	100.0	54.3
3	2	10	154 457	0.0	90.9	14	18	0	100.0	51.4
4	8	7	2375	0.3	63.6	23	11	0	100.0	31.4
5	9	7	2312	0.3	63.6	42	9	0	100.0	25.7
6	17	5	18	21.7	45.5	49	6	0	100.0	17.1
7	23	5	2	71.4	45.5	58	6	0	100.0	17.1
8	28	4	1	80.0	36.4	69	3	0	100.0	8.6
9	34	4	0	100.0	36.4	80	1	0	100.0	2.9
10	45	2	0	100.0	18.2	83	1	0	100.0	2.9
11	56	1	0	100.0	9.1	91	1	0	100.0	2.9
12	68	1	0	100.0	9.1	103	1	0	100.0	2.9
13	84	1	0	100.0	9.1	118	1	0	100.0	2.9
14	98	1	0	100.0	9.1	123	1	0	100.0	2.9
15	117	0	0	0.0	0.0	132	1	0	100.0	2.9
16	127	0	0	0.0	0.0	138	1	0	100.0	2.9
17	137	0	0	0.0	0.0	142	1	0	100.0	2.9
18	150	0	0	0.0	0.0	154	1	0	100.0	2.9
19	155	0	0	0.0	0.0	155	1	0	100.0	2.9

DMC

	5HT					ANA				
DEL	DS	ADC	DC	HR[%]	RR[%]	DS	ADC	DC	HR[%]	RR[%]
0	39	6	738	0.8	54.5	81	15	1	93.8	42.9
1	58	1	51	1.9	9.1	102	1	0	100.0	2.9
2	77	1	1	50.0	9.1	118	0	0	0.0	0.0
3	98	0	0	0.0	0.0	128	0	0	0.0	0.0
4	138	0	0	0.0	0.0	145	0	0	0.0	0.0
5	155	0	0	0.0	0.0	155	0	0	0.0	0.0

[a] The table shows how the numbers of database compounds and potential hits are reduced during dimension extension steps that produce descriptor reference spaces of increasing dimensionality and resolution. Abbreviations: DEL, dimension extension level; DS, descriptors; ADC, active database compounds (potential hits); DC, database compounds (i.e., ZINC compounds, considered false positives); HR, hit rate; RR, recovery rate.

other database compounds, 738 for 5HT and only 1 for ANA. The latter class represents an extreme case for the application of DMC because mapping the initial consensus position (81-bit signature) eliminates all but one ZINC compound and retains 15 hits. Thus, in some cases, DMC does not require dimension extension at all. Furthermore, the DMC results for 5HT illustrate that dimension extension typically leads to a sharp decline in compound numbers, which often results in small compound sets after a first or second step. Thus, although DMC utilizes binary-transformed descriptors, compound mapping conditions are by design more stringent than for DynaMAD. However, a common feature of these algorithms is that they are capable of enriching active compounds within very small sets of database molecules.

Differences in the dimension extension characteristics between DynaMAD and DMC are further illustrated in Figure 3, which shows exemplary comparisons of hit and recovery rates during dimension extension. Here, every data point represents a different dimension extension layer and the average of 100 independent trials (with bait sets of different compositions). Data points with the highest recovery rates correspond to the initially defined consensus position (DMC) or dimension extension layer 0 (DynaMAD). The figure represents the entire spectrum of class phenotypes we observed during virtual screening trials. Figure 3a shows two classes that could not be well-treated with either DynaMAD or DMC, whereas Figure 3b represents classes where DMC failed but DynaMAD produced promising results. In Figure 3c, classes are shown that gave very good results with both DMC and DynaMAD and displayed a similar path during dimension extension. By contrast, Figure 3d gives examples of classes where both methods also performed well but where significant differences observed in hit and recovery rates and their changes during dimension extension. The results in Figure 3 reflect intrinsic methodological differences between DynaMAD and DMC. For example, the profiles of ANA in Figure 3c or TKE in Figure 3d very well illustrate that changes in hit and recovery rates during dimension extension are typically more subtle and continuous for DynaMAD than for DMC. On the other hand, if DMC calculations succeed, they generally do so during the first one or two dimension extension steps where substantial changes in compound mapping occur.

5.6. Hit Rate Diagnostics. For "real-life" virtual screening applications, hit rates are usually more relevant as a performance measure than recovery rates because the total number of potential database hits is unknown. Thus, while Table 5 presents the results for the best combined hit and recovery rate (HRR), Table 7 reports the best hit rates achieved in our DynaMAD and DMC calculations (and also the corresponding dimension extension levels and recovery rates). Similar to the HRR results, DynaMAD generated higher hit rates than DMC for 20 of the 24 classes. These included three cases (5HT, BEN, and CAE) where DynaMAD produced hit rates greater than 50% or 60% but where DMC (on average) failed. In 19 cases, DynaMAD produced hit rates greater than 50%, and DMC did so in 11. In contrast to DMC, dimension extension levels producing top hit rates significantly varied for DynaMAD.

Optimizing hit rates in virtual screening generally lowers recovery rates because more stringent selection criteria need to be applied in order to minimize false-positive rates. As a consequence, compounds having similar activity but containing different structural motifs are also increasingly distinguished from each other.[4] Such trends are also seen in our calculations. When focusing on optimum hit rates, DynaMAD produced low recovery rates of less than 10% for nine classes and DMC produced low recovery rates for 12 classes. However, in a number of cases, recovery rates remained high.

Hit and recovery rates alone are not a reliable indicator of success in virtual screening calculations. The ability to correctly recognize different types of similarity relationships is another valid performance measure. Table 8 reports the number of hits that are shared by DynaMAD and DMC when focusing on the best single virtual screening trial yielding a selection set of maximal 20 compounds. No sets of correctly identified hits were identical, but there was overlap in 22 cases. For 10 activity classes, the smaller hit set (mostly produced with DMC) was completely contained in the larger one. However, for the remaining 14 classes, both DynaMAD and DMC produced unique hits not identified with the other method. These findings are well in accord with the known compound class dependence of virtual screening methods[47] and, importantly, demonstrate that DynaMAD and DMC are complementary in their use to increase the probability of identifying diverse hits. Figure 4 shows hits belonging to two activity classes that were correctly identified with DynaMAD and DMC and structures of representative bait compounds. The representation illustrates the structural diversity of compounds within activity classes and the ability of our mapping algorithms to detect remote similarity relationships.

5.7. Summary. In this study, we have introduced a novel compound mapping algorithm and evaluated its performance in comparison with two other mapping methods from which it was derived. DynaMAD utilizes combinations of compound-class-specific descriptor contributions to map compounds in chemical spaces of increasing dimensionality. The previously developed MAD approach laid the foundation for the identification and utilization of activity-class-specific or -selective descriptor settings. On the other hand, DMC made it possible to progressively map compounds to consensus positions in simplified descriptor spaces. DynaMAD combines the dynamic mapping features of DMC with the descriptor class value range analysis of MAD and utilizes further refined functions for descriptor identification, extension of class value ranges, and chemical space generation. Importantly, DynaMAD explores distributions of unmodified descriptor values for the comparison of active and inactive compounds. In extensive virtual screening trials on many different compound activity classes, DynaMAD has produced encouraging hit and recovery rates, with very few exceptions.

Because mapping algorithms are methodologically distinct from conventional virtual screening approaches, as further commented on below, it is difficult to include other methods in direct comparisons. Therefore, we have focused herein on a detailed comparison of DynaMAD with MAD and DMC. For MAD, we have previously carried out a comparison with a 2D structural fragment-type fingerprints.[46] This was done because fingerprint searching also generates similarity-ranked lists of candidate compounds. In these calculations, MAD produced consistently higher recovery rates and significantly higher hit rates for four of six activity classes that were studied.[38] As we have shown here,

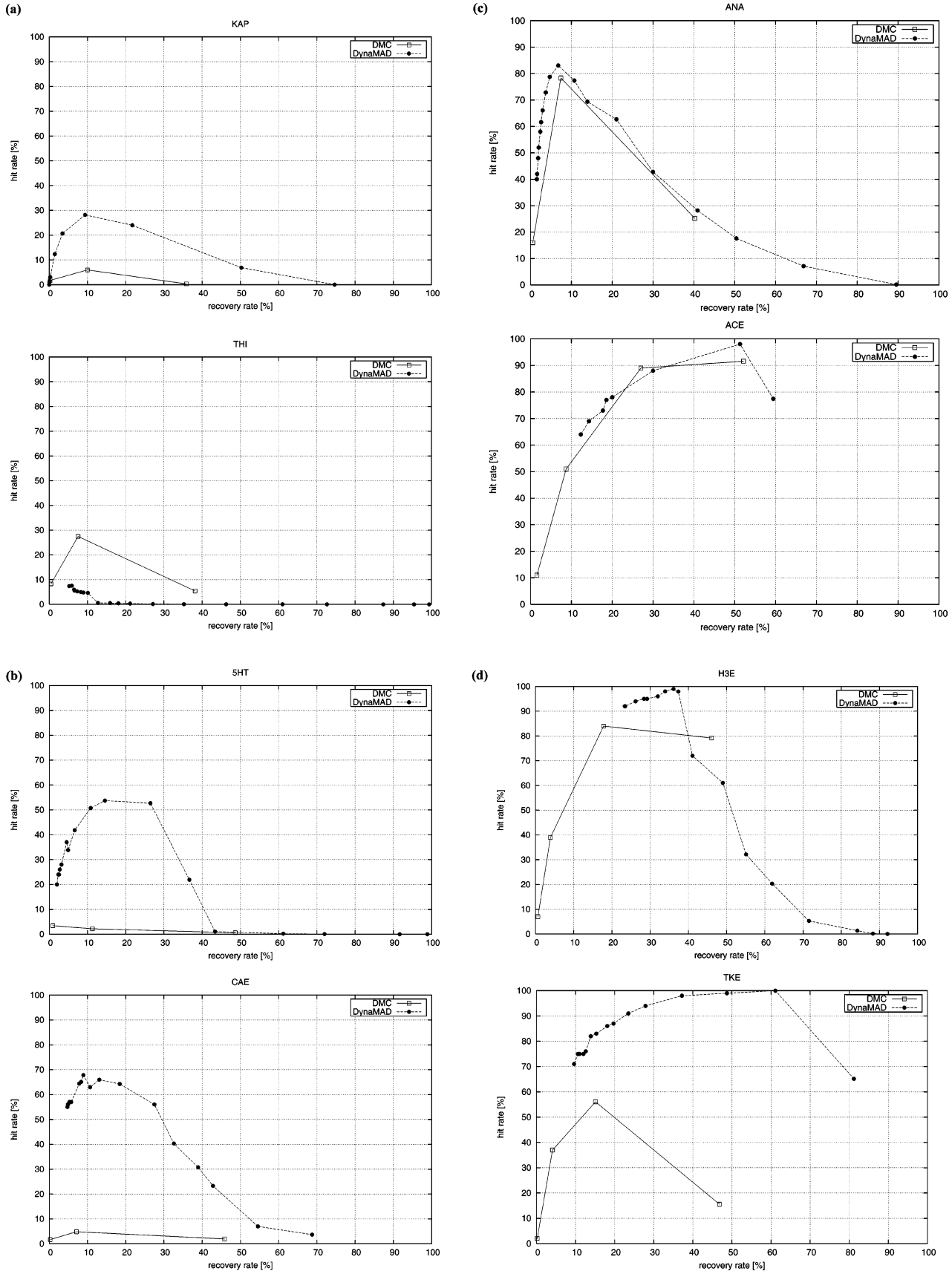

Figure 3. Comparison of hit and recovery rates during dimension extension. In a−d, representative activity class profiles are shown, as discussed in the text. Activity classes are designated according to Table 3. For both DMC and DynaMAD, each data point represents a different dimension extension level, averaged over 100 virtual screening trials. The dimensionality of the descriptor reference spaces increases in the order of decreasing recovery rates.

DynaMAD further improves the performance of both MAD and DMC, from which it was derived.

5.8. Concluding Remarks. Within the spectrum of currently available molecular similarity and virtual screening

Table 7. Top Hit Rates Per Class for DynaMAD and DMC[a]

activity class	DynaMAD			DMC		
	DEL	HR[%]	RR[%]	DEL	HR[%]	RR[%]
5HT	8	53.7	14.6	2	3.4	0.9
ACE	1	98.0	51.3	0	91.6	52.1
ANA	8	83.1	6.8	1	78.3	7.4
BEN	14	65.3	13.0	2	3.9	1.1
CAE	9	67.8	8.9	1	4.8	7.1
CHO	19	45.3	14.7	2	29.4	4.1
COX	0	100.0	97.3	1	63.1	18.6
DD1	17	94.7	16.0	1	64.2	14.3
DIR	2	57.2	26.5	2	7.4	1.5
EDN	8	32.9	14.7	1	48.3	4.6
ESU	2	60.2	40.3	1	44.1	11.4
GLY	10	90.7	14.3	1	86.5	10.4
H3E	9	99.0	36.1	1	84.0	17.8
HIV	1	100.0	44.4	1	68.0	23.0
INO	4	37.7	8.6	2	30.5	1.9
KAP	3	28.2	9.4	1	6.0	10.0
LAC	17	78.5	5.7	1	30.5	11.2
LDL	18	59.0	7.8	2	51.9	5.0
MEL	1	73.3	23.9	1	77.3	10.4
SQS	18	60.2	3.5	2	61.5	5.5
THI	18	7.6	5.9	1	27.5	7.5
THR	2	61.7	16.9	1	47.7	11.2
TKE	1	100.0	61.1	1	56.1	15.1
XAN	9	64.1	6.2	2	23.4	1.3

[a] The best hit rates for each activity class are reported, and the corresponding dimension extension levels and recovery rates are as well.

Table 8. Overlap in Hit Sets Identified with DynaMAD and DMC[a]

activity class	ADC	hits DynaMAD	hits DMC	common hits
5HT	11	5	1	1
ACE	7	5	6	5
ANA	35	19	15	11
BEN	12	7	2	1
CAE	12	6	1	0
CHO	20	6	6	5
COX	7	7	4	4
DD1	20	9	5	5
DIR	20	10	3	1
EDN	22	6	5	5
ESU	25	14	9	9
GLY	24	15	16	10
H3E	11	9	7	7
HIV	8	6	4	4
INO	25	8	3	2
KAP	15	4	2	0
LAC	19	6	6	4
LDL	20	5	8	4
MEL	15	6	6	3
SQS	32	11	9	6
THI	24	5	9	4
THR	23	18	5	5
TKE	10	10	6	6
XAN	25	8	12	7

[a] For each class, hits are reported for the best single trial producing a selection set of at most 20 compounds. ADC (active database compounds) reports the total number of potential hits added to ZINC. The number of hits identified with either DynaMAD or DMC is reported, and "common hits" gives the number of active compounds identified with both methods (selection overlap).

methodologies, mapping algorithms as described herein are characterized by a number of unique features and are only distantly related to (yet distinct from) other approaches such as statistical[25] or cell-based partitioning.[25,48] The conceptual relationship between mapping algorithms and cell-based partitioning methods lies in the fact that these approaches make use of independent descriptor contributions to distribute and position compounds in chemical reference spaces and do not rely on pairwise distance comparisons. By contrast, a fundamental difference is that DynaMAD does not require

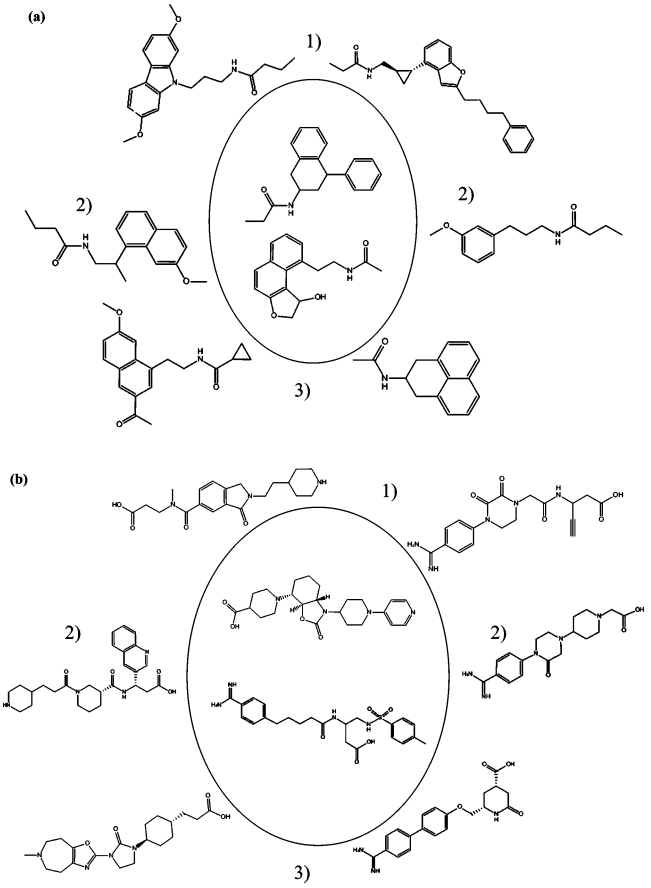

Figure 4. Diverse hits identified in virtual screening trials. For two activity classes, (a) MEL and (b) GLY, examples of hits are shown that were correctly identified with (1) only DynaMAD (two structures at the top), (2) DynaMAD and DMC (two structures in the middle), or (3) only DMC (two structures at the bottom). Circles in the center contain bait compounds used for the identification of these hits.

low-dimensional space representations[48] but operates in chemical spaces of increasing dimensionality. However, an interesting finding has been that DynaMAD is, at least in some instances, capable of accurately detecting remote similarity relationships in reference spaces formed by fewer than 10 descriptors, which sets it apart from DMC. This is a direct consequence of its ability to identify descriptors that specifically respond to compound activity classes and to use combinations of their value ranges for mapping. Irrespective of methodological details, our performance evaluation suggests that DynaMAD has significant potential for virtual screening applications.

ACKNOWLEDGMENT

We thank Martin Vogt for help in the assembly of compound classes and Jeffrey W. Godden for critical review of the manuscript.

NOTE ADDED IN PROOF

The mapping probability-based descriptor scoring function reported herein has now also been implemented in MAD and, as to be expected, further increased its virtual screening performance.

REFERENCES AND NOTES

(1) Brooijmans, N.; Kuntz, I. D. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biolmol. Struct.* **2003**, *32*, 335−373.

(2) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862−865.
(3) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Structure-based virtual screening and lead optimization: Methods and applications. *Nat. Rev. Drug Discovery* **2004**, *3*, 935−949.
(4) Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882−894.
(5) Green, D. V. Virtual screening of virtual libraries. *Prog. Med. Chem.* **2003**, *41*, 61−97.
(6) Stahura, F. L.; Bajorath, J. New methodologies for ligand-based virtual screening. *Curr. Pharm. Des.* **2005**, *11*, 1189−1202.
(7) Johnson, M.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, 1990.
(8) Martin, Y. C. 3D database searching in drug design. *J. Med. Chem.* **1992**, *35*, 2145−2154.
(9) Esposito, E. X.; Hopfinger, A. J.; Madura, J. D. Methods for applying the quantitative structure−activity relationship paradigm. *Methods Mol. Biol.* **2004**, *275*, 131−214.
(10) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(11) Barnard, J. M.; Downs, G. M. Chemical fragment generation and clustering software. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 141−142.
(12) Mason, J. S.; Cheney, D. L. Library design and virtual screening using multiple point pharmacophore fingerprints. *Pac. Symp. Biocomput.* **2000**, *5*, 576−587.
(13) Xue, L.; Godden, J. W.; Bajorath, J. Mini-fingerprints for virtual screening: Design principles and generation of novel prototypes based on information theory. *SAR QSAR Environ. Res.* **2003**, *14*, 27−40.
(14) Merlot, C.; Domine, D.; Cleva, C.; Church, D. J. Chemical substructures in drug discovery. *Drug Discovery Today* **2003**, *8*, 594−602.
(15) Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity searching using reduced graphs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 338−345.
(16) Putta, S.; Lemmen, L.; Beroza, P.; Greene, J. A novel shape-feature based approach to virtual library screening. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1230−1240.
(17) Brown, R. D.; Martin, Y. C. Use of structure−activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36* (6), 572−584.
(18) Engels, M. F. M.; Venkatarangan, P. Smart screening: Approaches to efficient HTS. *Curr. Opin. Drug Discovery Dev.* **2001**, *4*, 275−283.
(19) Tamura, S. Y.; Bacha, P. A.; Gruver, H. S.; Nutt, R. F. Data analysis of high-throughput screening results: Application of multi-domain clustering to the NCI anti-HIV data set. *J. Med. Chem.* **2002**, *45*, 3082−3093.
(20) Feher, M.; Schmidt, J. M. Fuzzy clustering as a means of selecting representative conformers and molecular alignments. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 810−818.
(21) Rusinko, A., III; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of a large structure/biological activity data set using recursive partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017−1026.
(22) Godden, J. W.; Furr, J. R.; Bajorath, J. Recursive median partitioning for virtual screening of large databases. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 182−188.
(23) Pearlman, R. S.; Smith, K. M. Novel software tools for chemical diversity. *Perspect. Drug Discovery Des.* **1998**, *9*, 339−353.
(24) Xue, L.; Godden, J.; Gao, H.; Bajorath, J. Identification of a preferred set of molecular descriptors for compound classification based on principal component analysis. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 699−704.
(25) Stahura, F. L.; Bajorath, J. Partitioning methods for the identification of active molecules. *Curr. Med. Chem.* **2003**, *8*, 707−715.
(26) Keseru, G. M.; Molnar, L.; Greiner, I. A neural network based virtual high throughput screening test for the prediction of CNS activity. *Comb. Chem. High Throughput Screening* **2000**, *3*, 535−540.
(27) Jorissen, R. N.; Gilson, M. K. Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.* **2005**, *44*, 549−561.
(28) Harper, G.; Bradshaw, J.; Gittin, J. C.; Green, D. V. S.; Leach, A. R. Prediction of biological activity for high-throughput screening using binary kernel discrimination. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1295−1300.
(29) Todeschini, R.; Consonni, V. Handbook of Molecular Descriptors. In *Methods and Principles in Medicinal Chemistry*; Mannhold, R., Kubinyi, H., Timmerman, H., Eds.; Wiley: New York, 2000; Vol. 11.
(30) Livingstone, D. J. The characterization of chemical structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195−209.
(31) Bajorath, J. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233−245.
(32) Agrafiotis, D. K.; Lobanov, V. S.; Salemme, R. F. Combinatorial informatics in the post-genomics era. *Nat. Drug Discovery Rev.* **2002**, *1*, 337−346.
(33) Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of biological activity profiles using substructural analysis and genetic algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165−179.
(34) Godden, J. W.; Bajorath, J. An information-theoretic approach to descriptor selection for database profiling and QSAR modeling. *QSAR Comb. Sci.* **2003**, *22*, 487−497.
(35) Hopfinger, A. J.; Reaka, A.; Venkatarangan, P.; Duca, J. S.; Wang, S. Construction of a virtual high throughput screen by 4D-QSAR analysis: Application to a combinatorial library of glucose inhibitors of glycogen phosphorylase b. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1151−1160.
(36) 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.
(37) Godden, J. W.; Stahura, F. L.; Bajorath, J. POT-DMC − A virtual screening method for the identification of potent hits. *J. Med. Chem.* **2004**, *47*, 4286−4290.
(38) 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.
(39) Godden, J. W.; Xue, L.; Kitchen, D. B.; Stahura, F. L.; Schermerhorn, E. J.; Bajorath, J. Median partitioning: A novel method for the selection of representative subsets from large compound pools. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 885−893.
(40) Cramer, R. D.; Jilek, R. J.; Guessregen, S.; Clark, S. J.; Wendt, B.; Clark, R. D. "Lead hopping". Validation of topomer similarity as a superior predictor of biological activities. *J. Med. Chem.* **2004**, *47*, 6777−6791.
(41) 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.
(42) *MOE (Molecular Operating Environment)*; Chemical Computing Group Inc.: Montreal, Quebec, Canada.
(43) Xue, L.; Bajorath, J. Accurate partitioning of compounds belonging to diverse activity classes. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 757−764.
(44) *Molecular Drug Data Report (MDDR)*; MDL Information Systems Inc.: San Leandro, CA.
(45) Rishton, G. M. Non-lead-likeness and lead-likeness in biochemical screening. *Drug Discovery Today* **2003**, *8*, 86−96.
(46) *MACCS Structural Keys*; MDL Information Systems Inc.: San Leandro, CA.
(47) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discovery Today* **2002**, *7*, 903−911.
(48) Pearlman, R. S.; Smith, K. M. Metric validation and the receptor-relevant subspace concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.

CI060083O