Bayesian Interpretation of a Distance Function for Navigating High-Dimensional Descriptor Spaces

Martin Vogt, Jeffrey W. Godden, and Jürgen Bajorath*

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

Received July 4, 2006

A distance function to analyze molecular similarity relationships in high-dimensional descriptor spaces and focus search calculations on "active subspaces" is defined in Bayesian terms. As a measure of similarity, database compounds are ranked according to their distance from the center of a subspace formed by known active molecules. From a Bayesian point of view, distance calculations are transformed into a "log-odds" estimate. Following this approach, maximizing the likelihood of a compound to be active corresponds to minimizing the distance from the center of an active subspace. Since the methodology generates a ranking of database molecules according to decreasing similarity to template compounds, it can be conveniently compared to similarity search tools, and the Bayesian function is found to compare favorably to two standard fingerprints in multiple template-based database searching.

INTRODUCTION

The design of chemical reference spaces using descriptors of molecular structures and properties is a prerequisite for the application of many computational methods to search for active molecules or design compound libraries.[1−3] Molecules are projected into multidimensional space representations based on their descriptor "coordinates" in order to, for example, classify them, estimate their chemical diversity, or analyze similarity relationships.[2,3] For many applications such as cell-based partitioning,[4] low-dimensional descriptor space representations are preferred,[4,5] and various methods have been developed or adapted for dimension reduction and orthogonalization of descriptor spaces[6−8] or de novo design of low-dimensional space representations.[3,4] However, it has also been shown that low-dimensional spaces are not essential for molecular similarity or diversity analysis and the detection of active compounds.[9−12] Therefore, we have investigated the development of conceptually different approaches to navigate high-dimensional space representations. These include partitioning methods that utilize binary-transformed property descriptors for the generation of simplified high-dimensional reference spaces,[13,14] activity-selective mapping algorithms that make use of either simplified[15] or original descriptor formulations,[16] and a function that uses Euclidean-like distances in high-dimensional space representation as a measure of molecular similarity.[17] This distance function was termed Distance in Activity-Centered Chemical Space (DACCS). It determines the distance of each database compound from the center of a region in chemical space that is populated by molecules having similar activity and produces a distance-based ranking of database compounds. Database molecules falling into the activity radius of a compound class were found to have a high probability to be active.[17] The simplicity and predictive ability of the DACCS function to capture molecular similarity relationships in complex descriptor spaces has encouraged us to investigate a methodologically distinct approach. Therefore, we have transformed the distance function into a Bayesian formulation that takes descriptor value distributions of active compounds and database molecules into account and establishes a novel relationship between distances in high-dimensional descriptor space and likelihood ratios. The approach has been evaluated on a large number of compound activity classes. In these calculations, we have observed a notable increase in the recovery of active compounds by the Bayesian function relative to the DACCS method and a favorable performance compared to well-known 2D fingerprints.

THEORY AND METHODS

We begin with an introduction of the original distance function. The DACCS function was designed to produce a distance-based ranking of database compounds in unmodified high-dimensional descriptor spaces.[17] As a measure of similarity, an Euclidean-like distance to the center of a subspace populated by a class of active compounds was calculated for each database molecule. If we operate based on the premise that similar molecules have similar activity, then increasing distance from the center of a subspace correlates with decreasing probability that a database compound is active (but no assumptions can a priori be made where a distance cutoff for activity might be). In order to alleviate the task of descriptor selection for the design of chemical reference spaces, we intended the approach to be applicable to chemical space representations generated by any number of arbitrarily chosen descriptors. Specifically, this was accomplished by application of a scaling procedure that centers chemical space on a subspace (defined by descriptor values of active compounds) and generates an orthogonal reference frame. The distance in scaled chemical space (d_{DACCS}) from the center of the "active subspace" to

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

10.1021/ci600280b CCC: $37.00
Published on Web 12/02/2006

each database compound is calculated as follows

$$d_{\mathrm{DACCS}} = \sqrt{\sum_{i=1}^{n}\left(\frac{x_i - \mu_i}{\sigma_i}\right)^2}$$

Here x_i is the value of descriptor i for compound x, μ_i is the mean value of descriptor i for a set of active compounds, and σ_i is the standard deviation of the descriptor values for these compounds. If this standard deviation is zero, then the standard deviation of the entire compound population is used instead to give such descriptors a unit weighting relative to the database compounds. Standard deviations of active compounds are expected to differ from the database if descriptors respond to activity class-specific features. In addition, sets of template compounds are only a relatively small sample giving rise to higher uncertainty in estimating the standard deviation. Both standard deviations can only be zero if a descriptor has the same value for all active and database compounds. In this special case, the descriptor has no information content and is omitted from the calculations. Importantly, DACCS calculations take all descriptor distributions into account and make no assumptions which might be most important.

Any high-dimensional descriptor space representation is affected by the so-called "curse of dimensionality",[18] which refers to the exponential growth of the volume of space as its dimensionality increases. As a consequence, a growing proportion of the chemical space falls outside of any hypersphere and into the corners of the enclosing hypercube. For database mining, this "curse" might turn into a relative 'blessing", in particular, when compound sets become very large, since the effects associated with increasing dimensionality can substantially reduce the number of database compounds that are located proximal to an "active subspace". This might explain why methods of little complexity like DACCS can successfully navigate high-dimensional space representations. In initial search calculations on five compound activity classes in a background database containing more than a million compounds, DACCS produced promising results. For small selection sets consisting of 25 database compounds, average recovery rates of up to ~88% were achieved.[17]

In the next step, we investigate the distance-based approach from a Bayesian point of view[19] and introduce a Bayesian approximation leading to a "log-odds" ratio estimation of activity. Given a set of known active molecules used as search templates (or "baits") and a background database, the likelihood that a compound belongs to an activity class A, given that descriptor i has value x_i, is calculated as follows:

$$L(A|x_i) \propto p(x_i|A) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right)$$

Here μ_i and σ_i are the mean and standard deviation of descriptor i for the given set of active molecules.

The central assumption in this approach is that the descriptor values follow a Gaussian distribution. Furthermore, if we also assume independence of the n descriptors, the likelihood of a compound with descriptor values $x = (x_i)_i$ to belong to activity class A is given as

$$L(A|x) \propto \prod_{i=1}^{n} p(x_i|\mathrm{A}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right)$$

If we then introduce a negative log-likelihood, we obtain the DACCS distance function:

$$-\log L(A|x) \propto \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{2\sigma_i^2}$$

The key aspect of this formulation is that *maximizing the likelihood corresponds to minimizing the distance* of descriptor values x to the activity center μ over all molecules in the database.

The approach can be extended by considering the likelihood ratio R (or odds) of a compound to belong to activity class A or to background database B. For this calculation, we need to determine the mean and standard deviation (ν_i and τ_i, respectively) of descriptor values for inactive database compounds. Since the background database also contains active compounds, this is an approximation. However, since the number of active molecules is typically very small compared to the number of database compounds, the statistical bias is negligible.

$$R(x) = \frac{L(A|x)}{L(B|x)} = \prod_{i=1}^{n} \frac{p(x_i|A)}{p(x_i|B)} = \prod_{i=1}^{n} \frac{\tau_i}{\sigma_i} \exp\left(\frac{(x_i - \nu_i)^2}{2\tau_i^2} - \frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right)$$

The ratio $R(x)$ will be greater than one if the probability density for active molecules is higher than the probability density for inactive molecules for the assumed Gaussian model, which is schematically illustrated in Figure 1. This means that the ratio $R(x)$ defines a subspace of descriptor values in which a compound has a higher probability of being active.

Taking the negative logarithm of $R(x)$ produces a *log-odds* value so that maximizing the odds is equivalent to minimizing the "distance"

$$-\log R(x) \propto \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{2\sigma_i^2} - \frac{(x_i - \nu_i)^2}{2\tau_i^2}$$

In addition to considering active molecules, the log-odds formulation explicitly takes the descriptor value distributions of database compounds into account. Since the DACCS function has been reformulated on the basis of Bayesian principles, we refer to the log-odds implementation as BDACCS.

CALCULATIONS

For BDACCS and DACCS calculations, a pool of 141 1D and 2D descriptors implemented in the Molecular Operating Environment (MOE)[20] was used. Reference calculations were carried out using two representative 2D fingerprints that differ in their design: a MOE-based

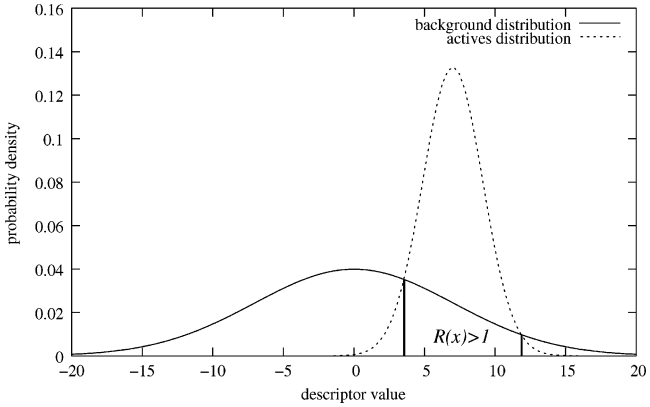

Figure 1. Descriptor distributions. This schematic representation shows value ranges for a hypothetical descriptor for database compounds (solid line) and active molecules (dashed). Typically, the descriptor distribution of a set of active compounds shows less variance than the value distribution in the background database. Modeling the distributions as Gaussians yields a range around the mean of the active distribution (i.e., the activity center) that is defined by the intersection of the curves and where the ratio of the distributions $R(x)$ is greater than 1. Active compounds are more likely to adopt values within this range than inactive molecules.

implementation of the well-known atom pair descriptors[21] in a 2-point pharmacophore-type fingerprint[20] (TGD, consisting of up to 735 bits when bond distances are permitted up to a maximum of 15 bonds) and a structural fragment fingerprint (MACCS, consisting of 166 bits) that encodes a set of publicly available MDL structural keys[22] (representing 166 substructures consisting of 1−10 non-hydrogen atoms). BDACCS and reference calculations were carried out in an in-house generated "2D unique" version of ZINC[23] containing ∼1.44 million compounds and using a total of 52 compound activity classes assembled from various databases[24−28] and the literature,[29−31] as summarized in Table 1. For similarity calculations, each activity class was individually added to the ZINC background database. For each class, 100 sets of 10 compounds each were randomly selected as templates for 100 independent search trials, and the remaining active molecules were added to ZINC. Thus, depending on the activity class, between 4 and 146 active molecules were available as potential hits within more than 1.4 million background molecules. Fingerprint search calculations using multiple template compounds were carried out after calculation of centroid fingerprints,[32] which resemble BDACCS calculation conditions in the sense that mean fingerprint settings are determined for a group of active compounds and utilized as the search template. Centroid fingerprint searches were found to increase similarity search performance relative to single template searches.[32] The corresponding individual fingerprints of database compounds were ranked by the Tanimoto coefficient (Tc)[33] relative to the centroid fingerprint. For each activity class and method, results were averaged over all search calculations.

All ZINC compounds were considered inactive, although the large source database is likely to contain hits against at least some of the 52 targets used here. Thus, the benchmark calculations provide a conservative low-end estimate of the virtual screening performance of the methodology because active database compounds are false-negatives in this analysis. One of the referees suggested attempting to predict whether top scoring ZINC compounds might also be active against the targets. Combinations of similarity search tools and similarity threshold values that can be reliably associated with the presence of biological activity are currently not available,[3,34,35] which precludes meaningful predictions; one could only speculate. For example, it has been demonstrated that a Tanimoto similarity threshold value of 0.85 in combination with Daylight fingerprints cannot be applied to

Table 1. Compound Activity Classes[a]

class designation	biological activity	no. of compds
	Activity Classes Assembled from the MDDR[24]	
AA2	adrenergic α-2 agonists	35
ACA	ACAT inhibitors	21
ANA	angiotensin II-AT antagonists	45
ARI	aldose reductase inhibitors	24
ARO	aromatase inhibitors	24
CAL	calpain inhibitors	28
CHO	cholesterol esterase inhibitors	30
DD1	dopamine D1 agonists	30
DIR	dihydrofolate reductase inhibitors	30
EDN	endothelin ETA antagonists	32
ESU	estrone sulfatase inhibitors	35
GLY	glycoprotein IIb-IIIa receptor antagonists	34
INO	inosine monophosphate dehydrogenase inhibitors	35
KAP	κ agonists	25
KRA	kainic acid receptor antagonists	22
LAC	lactamase β inhibitors	29
LDL	upregulator of LDL receptors	30
LIP	lipoxygenase inhibitors	41
MEL	melatonin agonists	25
PDE	phosphodiesterase type inhibitors	21
REN	renin inhibitors	51
SQS	inhibitors of squalene synthetase	42
THB	thrombin inhibitors	35
THI	thiol protease inhibitors	34
THR	thromboxane antagonists	33
XAN	xanthine oxidase inhibitors	35
	Activity Classes Assembled from the Literature[29−31]	
5HT	5-HT serotonin receptor ligands[29]	71
BEN	benzodiazepine receptor ligands[29]	59
BLC	β lactamase inhibitors[30]	14
CA	carbonic anhydrase II inhibitors[29]	159
CAA	calcium antagonists[30]	18
COX	cyclooxygenase-2 inhibitors[29]	31
D2A	dopamine D2 antagonists[30]	14
GHS	growth hormone secretagogue agonists[31]	14
GRH	gonadotropin releasing hormone agonists[31]	100
H3	H3 antagonists[29]	52
HIV	HIV protease inhibitors[29]	48
JNK	C-jun N-terminal kinase inhibitors[31]	36
MCH	melanin-concentrating hormone[31]	30
PAR	PPAR γ agonists[30]	16
PKC	protein kinase C inhibitors[30]	15
RTI	reverse transcriptase inhibitors[30]	15
TK	tyrosine kinase inhibitors[29]	35
	Activity Classes Assembled from Other Sources[25−28]	
ADR	β-receptor anti-adrenergics[25]	16
CDK1	cyclin-dependent kinase 1 inhibitors[27]	22
CDK2	cyclin-dependent kinase 2 inhibitors[27]	24
FAC	factor Xa inhibitors[28]	14
GLU	glucocorticoid analogues[25]	14
H1D	histamine H1 receptor antagonists[26]	36
M2	muscarinic M2 receptor antagonists[26]	20
NET	norepinephrine transporter inhibitors[26]	21
VEG	VEGFR-2 tyrosine kinase inhibitors[27]	36

[a] Compound activity classes were assembled from a variety of databases and the literature. For classes assembled from the MDDR, between approximately 20 and 50 compounds were randomly selected in each case to produce compound sets comparable in size to those obtained from other sources. MDDR compounds were only included in the activity classes used here if they had references and/or potency information associated with them.

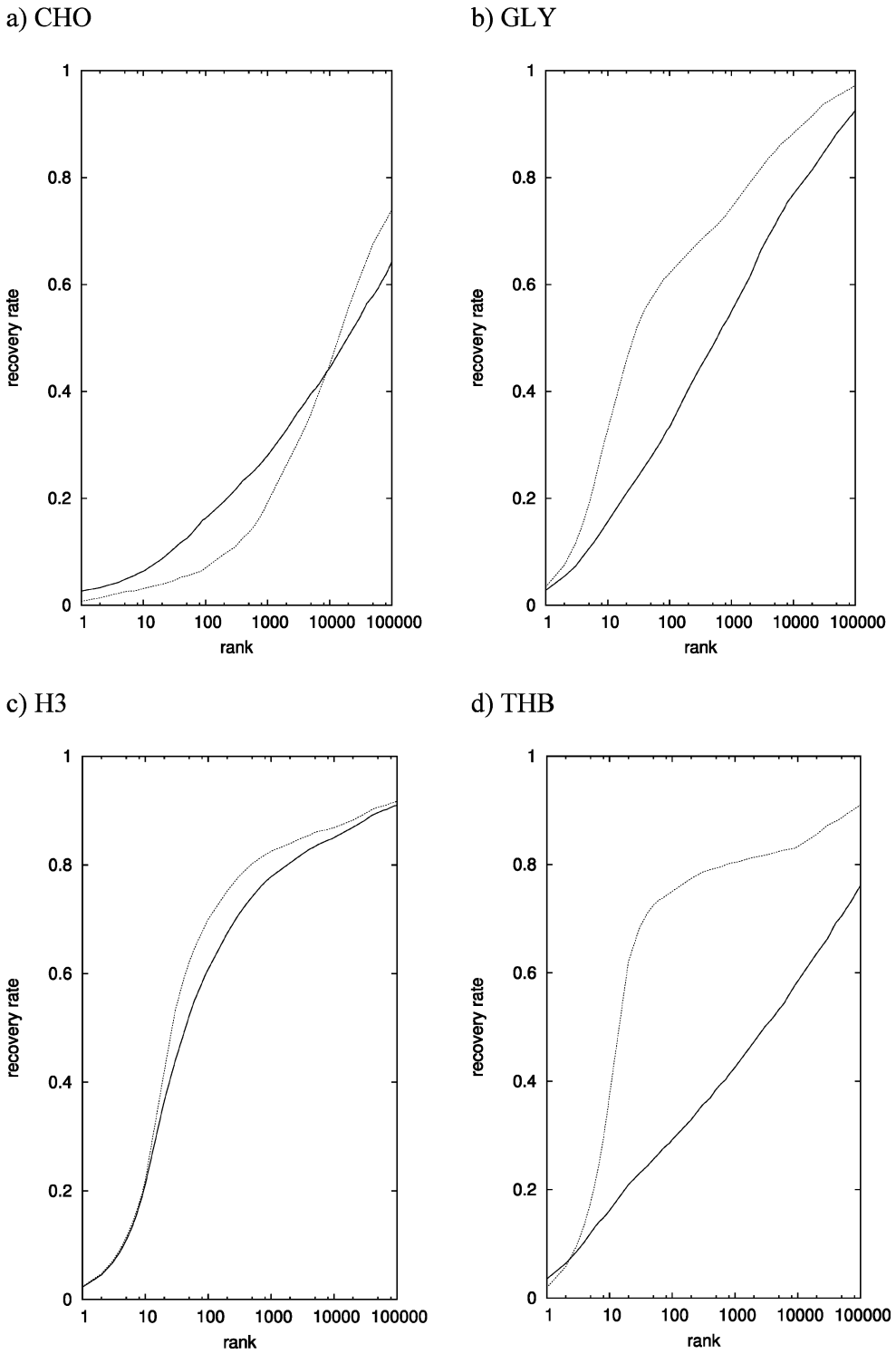

Figure 2. Comparison of DACCS and BDACCS. Shown are representative recall curves for active molecules. Recovery rates are reported for increasingly larger numbers of ranked database compounds obtained with DACCS (solid line) and BDACCS (dashed). For abbreviations, see Table 1.

reliably predict whether or not database compounds are active.[35] Thus, the potential activity of ZINC database compounds could not be confirmed without experimental evaluation, which is, however, not the purpose of our benchmark calculations. For our study, it is important to note that with 52 different compound sets a very large number of activity classes have been tested in order to evaluate and confirm the predictive ability of the BDACCS function.

RESULTS AND DISCUSSION

We first compared the ability of BDACCS and DACCS to recover active compounds from our source database. Recall curves representing the different performance ratios we observed in these calculations are shown in Figure 2 (and recall curves for all 52 activity classes are provided in the Supporting Information, Figure 1). We detected either slightly better recovery rates for DACCS (Figure 2a),

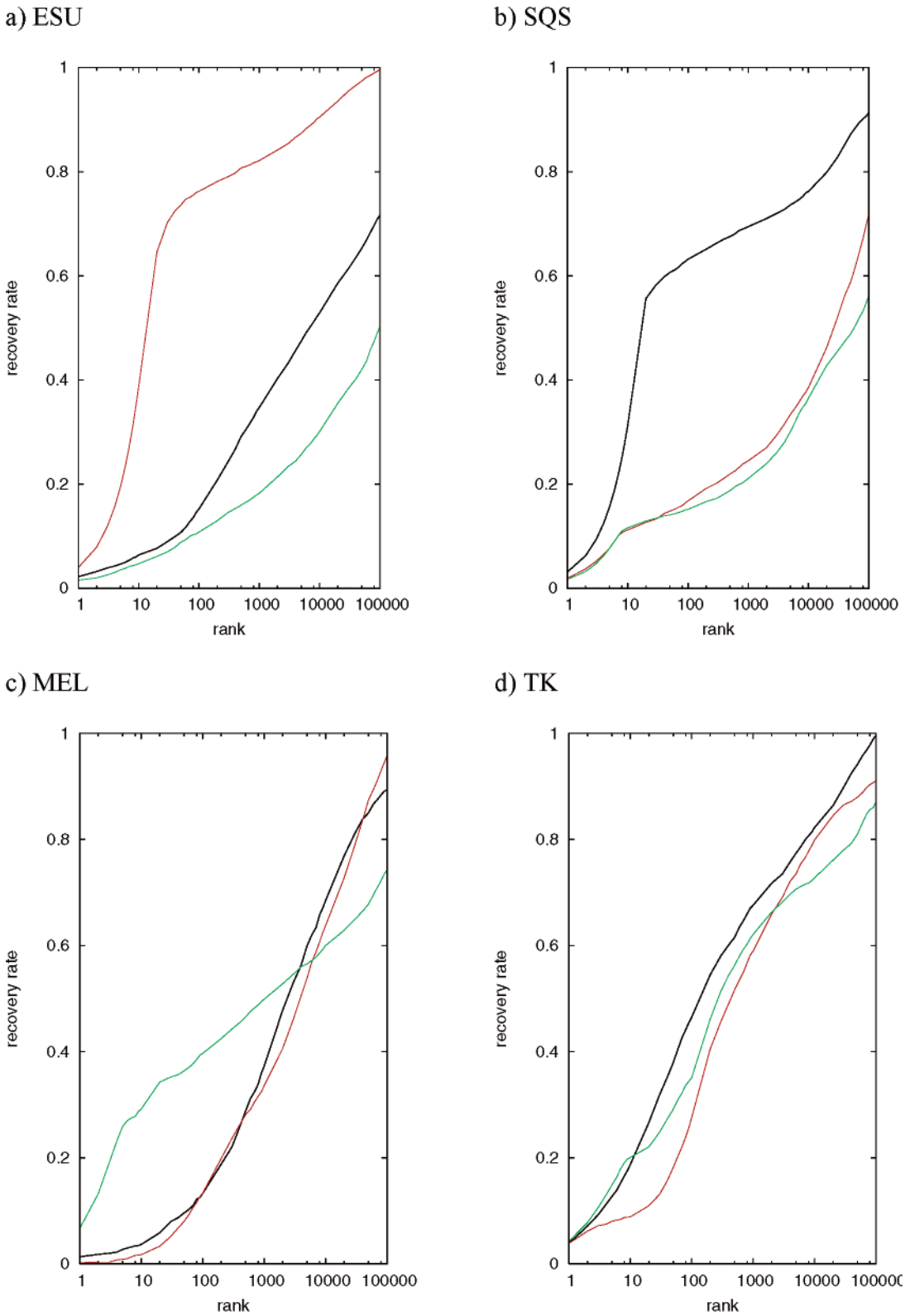

Figure 3. Comparison of BDACCS and similarity search calculations. Representative recall curves are presented according to Figure 2 for BDACCS (black), MACCS (red), and TGD (green). Corresponding recovery rates are reported in Table 2. For abbreviations, see Table 1.

comparable performance (Figure 2c), or moderately (Figure 2b) to significantly better (Figure 2d) recovery rates for BDACCS. Overall, BDACCS displayed a systematic improvement in the recovery of active compounds relative to DACCS. In 35 of 52 activity classes, BDACCS had overall higher recovery rates, and, in 10 of these classes, the improvements were rather significant. In one case, JNK (Table 1), DACCS calculations failed, whereas BDACCS produced recovery rates of ~23% and ~36% within the top-ranked 10 and 100 database compounds, respectively. For 15 other classes, BDACCS and DACCS produced comparable results, and, in only two cases, DACCS was superior.

Table 2. Compound Recovery Rates (100 Trials)[a]

	method										method								
	BDACCS mean % (SD)			MACCS mean % (SD)			TGD mean % (SD)				BDACCS mean % (SD)			MACCS mean % (SD)			TGD mean % (SD)		
	10[b]	100[b]	1000[b]	10[b]	100[b]	1000[b]	10[b]	100[b]	1000[b]		10[b]	100[b]	1000[b]	10[b]	100[b]	1000[b]	10[b]	100[b]	1000[b]
5HT	8.5	26.4	44.5	7.0	23.4	54.6	4.1	11.0	20.0	H1D	3.4	11.8	27.6	6.3	16.1	26.3	6.4	13.1	20.5
	(3.2)	(6.8)	(12.4)	(4.2)	(6.4)	(12.2)	(5.9)	(8.6)	(8.6)		(3.8)	(7.6)	(8.1)	(3.8)	(6.7)	(6.4)	(6.5)	(8.9)	(8.6)
AA2	0.2	1.6	9.3	3.4	7.9	18.9	0.8	2.2	7.0	H3	22.0	70.0	82.5	21.9	69.1	85.4	18.7	44.1	64.0
	(1.1)	(3.1)	(7.5)	(3.8)	(5.2)	(8.4)	(2.0)	(3.8)	(6.8)		(1.5)	(10.1)	(9.1)	(1.4)	(6.4)	(5.4)	(3.8)	(13.4)	(14.5)
ACA	19.0	21.9	29.4	0.1	1.5	6.7	0.0	0.0	0.0	HIV	26.3	87.6	92.9	26.3	85.7	88.8	23.5	47.3	61.0
	(9.0)	(9.7)	(11.6)	(0.9)	(3.7)	(7.3)	(0.0)	(0.0)	(0.0)		(0.0)	(3.7)	(3.2)	(0.0)	(5.8)	(3.0)	(4.5)	(11.1)	(10.6)
ADR	11.5	57.0	73.0	27.2	75.7	97.2	10.8	26.5	66.5	INO	10.1	20.6	33.2	9.9	20.8	43.4	13.0	27.6	37.5
	(12.0)	(14.4)	(13.8)	(15.7)	(14.1)	(6.3)	(11.2)	(17.6)	(16.2)		(8.3)	(10.5)	(12.6)	(7.8)	(9.8)	(9.2)	(5.2)	(8.3)	(7.1)
ANA	27.4	50.5	67.8	22.5	44.3	59.2	28.5	73.3	85.4	JNK	22.7	36.2	42.2	1.9	3.9	10.4	0.7	1.0	2.0
	(2.5)	(11.1)	(11.7)	(5.9)	(7.7)	(7.4)	(0.4)	(6.6)	(3.1)		(7.4)	(6.4)	(5.6)	(3.1)	(4.8)	(8.1)	(2.3)	(2.9)	(4.0)
ARI	3.1	6.4	14.8	0.0	1.1	5.3	0.0	0.1	1.1	KAP	5.3	13.1	30.2	8.4	11.4	20.9	9.4	12.3	22.3
	(4.2)	(6.0)	(8.3)	(0.0)	(2.6)	(5.3)	(0.0)	(0.7)	(2.8)		(5.0)	(7.5)	(11.6)	(5.6)	(5.7)	(7.6)	(11.8)	(12.8)	(11.3)
ARO	0.0	0.5	9.1	0.3	1.4	7.0	0.7	2.9	5.6	KRA	3.2	13.9	31.1	12.8	19.2	28.8	3.9	11.2	22.0
	(0.0)	(1.8)	(5.7)	(1.7)	(3.4)	(7.4)	(2.2)	(3.7)	(5.5)		(5.3)	(10.4)	(12.4)	(9.9)	(11.9)	(12.6)	(5.1)	(9.0)	(12.5)
BEN	14.0	32.0	50.9	17.0	58.0	88.0	7.3	13.6	21.8	LAC	5.8	16.2	36.1	11.9	17.3	24.9	3.2	5.7	12.3
	(4.9)	(6.8)	(10.3)	(4.1)	(16.5)	(8.7)	(6.4)	(9.9)	(10.4)		(6.6)	(10.8)	(12.5)	(7.7)	(9.6)	(11.5)	(4.0)	(5.2)	(7.5)
BLC	67.8	86.8	100.0	88.0	93.2	94.0	18.8	27.5	39.0	LDL	18.0	27.5	39.6	11.1	20.2	32.9	20.9	32.1	39.6
	(20.8)	(12.5)	(0.0)	(15.3)	(11.2)	(10.7)	(17.5)	(19.9)	(23.4)		(6.7)	(6.4)	(8.0)	(11.6)	(12.8)	(13.5)	(8.5)	(9.4)	(9.6)
CA	4.3	18.2	45.6	5.7	30.0	52.7	1.5	4.8	8.1	LIP	6.6	13.0	24.9	14.5	21.0	30.5	5.9	10.5	15.9
	(2.2)	(7.7)	(13.8)	(1.8)	(9.0)	(8.1)	(1.6)	(3.0)	(4.8)		(6.4)	(9.7)	(12.9)	(6.7)	(8.5)	(9.7)	(6.5)	(8.2)	(9.4)
CAA	24.8	41.4	64.8	16.2	40.0	69.5	0.2	5.1	17.5	M2	0.5	3.9	12.8	0.4	1.5	7.1	0.9	2.9	5.8
	(12.7)	(15.5)	(18.2)	(10.7)	(14.8)	(18.6)	(1.8)	(7.1)	(10.8)		(2.2)	(6.0)	(8.5)	(2.0)	(4.1)	(7.7)	(2.9)	(4.6)	(5.4)
CAL	7.2	13.7	23.1	7.5	12.7	20.2	2.2	5.3	9.6	MCH	23.2	35.9	45.6	0.1	1.6	8.3	1.9	6.4	19.9
	(4.7)	(7.6)	(8.8)	(8.1)	(10.9)	(12.3)	(3.4)	(4.8)	(6.6)		(7.5)	(7.9)	(8.3)	(0.7)	(2.5)	(5.3)	(3.2)	(6.8)	(9.1)
CDK1	26.8	45.9	62.0	29.7	40.8	51.6	7.4	9.2	12.9	MEL	3.7	13.3	37.2	1.7	13.3	33.5	29.2	39.7	49.9
	(10.9)	(13.3)	(12.6)	(12.7)	(13.4)	(13.3)	(8.6)	(9.7)	(10.8)		(4.8)	(7.6)	(9.3)	(3.4)	(6.6)	(9.4)	(7.8)	(8.4)	(10.2)
CDK2	26.8	43.0	58.3	35.2	45.4	57.5	3.1	3.6	5.6	NET	0.0	2.8	13.4	1.6	3.4	6.8	0.2	1.9	6.9
	(10.4)	(11.9)	(13.5)	(18.8)	(17.6)	(17.6)	(5.6)	(6.1)	(7.5)		(0.0)	(5.1)	(9.3)	(3.5)	(4.8)	(6.5)	(1.3)	(3.9)	(7.7)
CHO	3.1	7.0	19.2	14.2	36.4	49.2	0.3	1.4	3.5	PAR	15.3	24.0	33.7	31.2	41.7	69.3	21.8	31.7	40.3
	(6.1)	(10.0)	(13.7)	(5.7)	(11.4)	(12.6)	(1.2)	(2.7)	(4.0)		(11.8)	(14.9)	(14.8)	(14.1)	(16.7)	(18.0)	(13.5)	(14.5)	(15.0)
COX	44.7	80.2	91.6	32.8	52.3	70.0	21.4	35.0	51.4	PDE	3.1	4.3	7.3	0.0	0.3	2.4	0.0	0.2	1.0
	(4.3)	(8.1)	(7.7)	(10.7)	(14.2)	(10.9)	(8.4)	(13.0)	(12.3)		(5.0)	(5.6)	(7.0)	(0.0)	(1.6)	(4.9)	(0.0)	(1.3)	(2.9)
D2A	12.2	24.8	49.8	31.0	39.0	52.8	0.0	1.2	13.2	PKC	41.2	67.8	83.8	57.2	58.4	58.4	15.4	18.6	29.0
	(16.9)	(21.2)	(24.5)	(18.5)	(22.8)	(26.8)	(0.0)	(5.5)	(17.9)		(24.4)	(24.1)	(16.5)	(18.2)	(19.0)	(19.0)	(15.3)	(16.1)	(17.8)
DD1	18.7	46.5	67.6	8.8	27.6	59.0	20.1	35.1	62.1	RTI	0.0	0.4	53.0	3.2	23.2	63.6	21.0	49.0	56.2
	(8.0)	(10.2)	(9.7)	(5.7)	(9.4)	(12.8)	(9.4)	(10.8)	(10.6)		(0.0)	(2.8)	(15.7)	(7.9)	(17.0)	(20.0)	(16.2)	(12.8)	(15.5)
DIR	9.7	26.1	48.6	8.8	25.4	54.2	8.6	17.0	26.2	REN	24.4	95.3	95.4	23.7	64.1	79.2	20.7	60.9	75.0
	(8.2)	(13.0)	(14.0)	(6.2)	(7.8)	(9.0)	(5.2)	(8.4)	(9.2)		(0.0)	(8.4)	(8.3)	(1.5)	(12.0)	(11.0)	(4.0)	(10.3)	(9.0)
EDN	38.9	50.0	55.9	0.5	3.0	23.7	5.2	8.7	15.8	SQS	31.2	63.2	69.5	11.2	16.8	24.5	11.6	15.2	21.1
	(6.3)	(7.2)	(6.9)	(1.5)	(3.4)	(7.6)	(4.0)	(5.9)	(8.3)		(0.0)	(6.4)	(6.5)	(9.7)	(10.3)	(11.5)	(11.5)	(12.9)	(13.6)
ESU	6.4	15.2	34.7	38.7	76.2	82.2	4.7	10.8	18.2	THI	4.2	11.4	25.5	0.0	0.5	3.6	0.1	0.8	4.8
	(6.5)	(11.6)	(13.2)	(2.3)	(6.8)	(5.6)	(4.5)	(5.9)	(7.4)		(3.5)	(5.2)	(8.1)	(0.4)	(1.5)	(4.1)	(0.6)	(1.8)	(4.1)
FAC	48.5	61.3	68.8	27.5	53.5	72.2	1.8	3.8	5.2	THR	25.7	34.8	45.1	9.3	20.9	33.8	20.0	29.7	39.0
	(20.7)	(21.4)	(21.1)	(19.9)	(23.8)	(25.1)	(6.4)	(9.0)	(10.2)		(6.4)	(7.9)	(9.6)	(7.1)	(10.3)	(11.4)	(9.1)	(9.0)	(8.1)
GHS	84.2	86.0	90.5	12.2	24.8	34.0	0.8	11.8	22.2	TK	32.0	59.1	77.5	15.1	42.9	77.8	18.3	22.6	31.2
	(18.3)	(16.0)	(13.2)	(14.0)	(19.6)	(22.6)	(4.3)	(14.4)	(17.7)		(6.3)	(13.7)	(11.5)	(6.7)	(12.1)	(12.7)	(5.2)	(5.5)	(7.5)
GLU	17.2	61.3	87.8	25.5	59.2	84.8	32.8	77.2	78.2	THB	37.4	75.0	80.4	11.9	19.2	31.4	27.3	31.4	36.4
	(15.8)	(20.8)	(18.3)	(18.1)	(21.8)	(15.4)	(20.9)	(19.5)	(19.4)		(2.0)	(8.4)	(8.6)	(7.2)	(9.8)	(10.1)	(7.5)	(7.7)	(7.7)
GLY	32.8	62.1	74.4	7.4	16.8	37.8	27.1	36.6	50.7	VEG	3.8	16.4	42.2	7.3	18.2	44.6	1.8	4.2	10.6
	(7.4)	(9.4)	(11.1)	(4.4)	(7.2)	(9.8)	(8.7)	(11.8)	(10.4)		(4.4)	(8.9)	(10.9)	(6.7)	(9.0)	(10.4)	(2.8)	(4.3)	(6.2)
GRH	11.1	65.4	82.3	9.9	46.8	57.2	4.3	8.8	13.5	XAN	19.4	36.3	50.5	20.2	31.6	46.2	12.6	19.5	32.2
	(0.0)	(7.2)	(7.4)	(2.6)	(13.7)	(10.5)	(4.1)	(7.2)	(7.2)		(7.5)	(9.3)	(9.0)	(8.0)	(10.2)	(9.9)	(8.9)	(11.0)	(15.3)

[a] For each activity class, compound recovery rates are reported in percent as averages over 100 independent trials for differently sized compound selection sets, i.e., active molecules among the top 10, 100, or 1000 database compounds, ranked by log-odds value (BDACCS) or Tc (MACCS, TGD). Standard deviations (SD) are given in parentheses. [b] Selection set size.

Thus, BDACCS performed systematically better than DACCS in large-scale calculations.

In the next step, we compared BDACCS with multiple template-based similarity searching using two fingerprints encoding well-established descriptors, MACCS and TGD, that differ in their design. Prototypic recall curves are shown in Figure 3. Either one of the three methods performed best (Figure 3a−c), or the performance of at least two methods was found to be comparable (Figure 3d). Recall curves for all 52 activity classes are presented in the Supporting Information, Figure 2, and corresponding recovery rates for different numbers of top-ranked database compounds (10, 100, or 1000) are reported in Table 2. Recall rate standard deviations on the order of 20% were observed in a number of instances, which clearly demonstrates the need to carry out a significant number of individual calculations on each class (100 in this case) in order to balance the template set-dependence of these calculations. The comparison of BDACCS and fingerprint search calculations also revealed some systematic trends. The relative performance of these similarity methods displayed a strong compound class-dependence, which is a well-known phenomenon in similarity search-

ing[34,35] and ligand-based virtual screening.[3,12] In 17 of 52 cases, the performance of at least two of the methods was comparable. However, BDACCS produced overall best results on 22 activity classes, whereas MACCS and TGD performed best on 14 and 4 classes, respectively. For two of the activity classes (ARO, M2), all three methods essentially failed. For five other classes (AA2, ARI, CHO, NET, PDE), only low recovery rates were achieved that were comparable to their standard deviations. Thus, in these cases BDACCS calculations could not reliably recover active molecules among the top 1000 database compounds. In 14 other cases (ACA, CA, CAA, D2A, DIR, ESU, H1D, INO, KAP, KRA, LAC, LIP, PAR, VEG), not only significant recovery rates but also relatively high standard deviations were observed, thus indicating considerably dependence of the results on the composition of bait sets in independent trials. In some cases, very significant differences in recovery rates were observed, for example, for ESU (Figure 3a) or SQS (3b) where MACCS and BDACCS, respectively, outperformed the other methods. BDACCS also performed much better than the reference fingerprints on five other classes (ACA, EDN, JNK, MCH, THI) (Table 2). In one of these cases (ACA), MACCS and TGD failed to recover any active molecules, whereas BDACCS produced recovery rates of ∼19% and ∼22% for the top 10 and 100 database compounds, respectively.

The major goal of our test calculations has been to evaluate the potential of the BDACCS approach to recover active compounds from high-dimensional descriptor spaces on the basis of log-odds scores. The results of our large-scale evaluation provided evidence that both DACCS and BDACCS are capable of successfully detecting molecular similarity relationships for many different classes of active compounds added to a large background database. Furthermore, our findings confirmed a systematic improvement of the BDACCS approach relative to DACCS from which it was derived. This performance improvement is attributed to the fact that BDACCS also takes the descriptor value distributions of database compounds into account, in addition to defining an activity class-oriented subspace. Since BDACCS produces a probability-based ranking of database compounds, it can be compared to similarity search tools such as fingerprints that also rank compounds based on similarity metrics. For these purposes, we have selected two 2D fingerprints that encode a magnitude of descriptor contributions comparable to the descriptor pool used here. In addition, we have applied a multiple template-based search strategy that has previously been shown to produce encouraging results[32] and that is also reminiscent of the activity centering scheme of BDACCS. The comparison of recovery rates produced by our log-odds calculations and centroid fingerprint searches revealed that BDACCS compared overall favorably to these standard search tools. In addition, it performed well in a number of cases where compound recovery using fingerprints was low. Thus, these findings indicate that BDACCS complements currently available molecular similarity methods. An attractive feature of the BDACCS approach is that it can operate in any descriptor space, irrespective of its dimensionality or composition. In our calculations, we have not distinguished between potency levels of active compounds, but potency values could in principle be used to scale compound contributions when calculating the center of an active subspace to more accurately map its position. Conceptually similar potency scaling procedures were previously applied to increase the probability of detecting potent hits in virtual screening calculations using a mapping algorithm.[36]

CONCLUSIONS

We have introduced a Bayesian model to navigate high-dimensional descriptors spaces that transforms a scaled Euclidean distance function into a probabilistic function and further increases recovery rates of molecules belonging to many different activity classes. DACCS and BDACCS are methodologically distinct from other similarity-based methods and alleviate the need for descriptor selection and space design because they can detect similarity relationships in high-dimensional and unrefined references spaces. Given the results of our test calculations, we conclude that BDACCS further adds to the spectrum of contemporary similarity methods. Its computational efficiency and effectiveness on diverse activity classes suggest that BDACCS calculations should have considerable potential for virtual screening.

Supporting Information Available: For all 52 activity classes, recall curves produced with BDACCS, DACCS, and two fingerprints (Figures 1 and 2). This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Bajorath, J. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233−245.
(2) Agrafiotis, D. K.; Lobanov, V. S.; Salemme, R. F. Combinatorial informatics in the post-genomics era. *Nature Drug Discovery Rev.* **2002**, *1*, 337−346.
(3) Bajorath, J. Integration of virtual and high-throughput screening. *Nature Rev. Drug Discovery* **2002**, *1*, 882−894.
(4) Pearlman, R. S.; Smith, K. M. Novel software tools for chemical diversity. *Perspect. Drug Discovery Des.* **1998**, *9*, 339−353.
(5) Pearlman, R. S.; Smith, K. M. Metric validation and the receptor-relevant subspace concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.
(6) Xue, L.; Bajorath, J. Molecular descriptors for effective classification of biologically active compounds based on principal component analysis identified by a genetic algorithm. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 801−809.
(7) Agrafiotis, D. K.; Lobanov, V. S. Nonlinear mapping networks. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1356−1362.
(8) Agrafiotis, D. K.; Lobanov, V. S. Multi-dimensional scaling of combinatorial libraries without explicit enumeration. *J. Comput. Chem.* **2001**, *22*, 1712−1722.
(9) Godden, J. W.; Xue, L.; Bajorath, J. Classification of biologically active compounds by median partitioning. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1263−1269.
(10) 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.
(11) Jorissen, R. N.; Gilson, M. K. Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.* **2005**, *44*, 549−561.
(12) Stahura, F. L.; Bajorath, J. New methodologies for ligand-based virtual screening. *Curr. Pharm. Des.* **2005**, *11*, 1189−1202.
(13) 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.
(14) Godden, J. W.; Bajorath, J. Partitioning in binary-transformed descriptor spaces. *Meth. Mol. Biol.* **2004**, *275*, 291−300.
(15) 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.
(16) 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.

(17) Godden, J. W.; Bajorath, J. A distance function for retrieval of active molecules from complex chemical space representations. *J. Chem. Inf. Model.* **2006**, *46*, 1094−1097.
(18) Bellman, R. E. *Adaptive Control Processes*; Princeton University Press: Princeton, NJ, 1961; pp 1−255.
(19) Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*, 2nd ed.; Wiley-Interscience: New York, 2000; pp 20−83.
(20) *Molecular Operating Environment (MOE), Vers. 2005.06*; Chemical Computing Group Inc., 1255 University Street, Montreal, Quebec, Canada, H3B 3X3. http://www.chemcomp.com (accessed June 2006).
(21) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical similarity using geometric atom pair descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128−136.
(22) *MACCS structural keys*; MDL Information Systems Inc.: San Leandro, CA, U.S.A., 2002.
(23) 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.
(24) *Molecular Drug Data Report (MDDR)*; MDL Information Systems Inc.: 14600 Catalina Street, San Leandro, CA, U.S.A., 2005.
(25) *Comprehensive Medicinal Chemistry Database (CMC-3D), Version 99.1*; MDL Information Systems Inc.: 14600 Catalina Street, San Leandro, CA, 1999.
(26) Roth, B. L.; Kroeze, W. K; Patel, S.; Lopez, E. The multiplicity of serotonin receptors: uselessly diverse molecules or an embarrassment of riches? *The Neuroscientist* **2000**, *6*, 252−262. http://kidb.bioc.cwru.edu (accessed June 2006).
(27) Chen, X.; Liu, M.; Gilson, M. K. Binding DB: a web-accessible molecular recognition database. *J. Comb. Chem. High Throughput Screening* **2001**, *4*, 719−725. http://www.bindingDB.org (accessed June 2006).
(28) Synthline Drug Database on STN International, taken from *Drugs of the Future* (comprehensive drug monographs, Prous Science), 1984−present; Prous Science: Provenza 388, Barcelona, Spain.
(29) Xue, L.; Bajorath, J. Accurate partitioning of compounds belonging to diverse activity classes. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 757−764.
(30) Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J. Profile scaling increases the similarity search performance of molecular fingerprints containing numerical descriptors and structural keys. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1218−1225.
(31) Godden, J. W.; Florence, F. L.; Bajorath, J. Anatomy of fingerprint search calculations on structurally diverse sets of active compounds. *J. Chem. Inf. Model.* **2005**, *45*, 1812−1819.
(32) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity metrics for ligands reflecting the similarity of the target protein. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391−405.
(33) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(34) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods. *Drug Discovery Today* **2002**, *7*, 903−911.
(35) Martin, Y. C.; Kofron, J. C.; Traphagen, L. M. Do structurally similar molecules have similar biological activity. *J. Med. Chem.* **2002**, *45*, 4350−4358.
(36) Godden, J. W.; Stahura, F. L.; Bajorath, J. POT-DMC: a virtual screening method for the identification of potent hits. *J. Med. Chem.* **2003**, *47*, 5608−5611.

CI600280B