ARTICLES

Bayesian Similarity Searching in High-Dimensional Descriptor Spaces Combined with Kullback-Leibler Descriptor Divergence Analysis

Martin Vogt and Jürgen Bajorath*

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

Received September 7, 2007

We investigate an approach that combines Bayesian modeling of probability distributions of descriptor values of active and database molecules with Kullback-Leibler analysis of the divergence between these distributions. The methodology is used for Bayesian screening and also to predict compound recall rates. In our study, we analyze two fundamental approximations underlying the Bayesian screening approach: the assumption that descriptors are independent of each other and, furthermore, that their data set values follow normal distributions. In addition, we calculate Kullback-Leibler divergence for single descriptors, rather than multiple-feature distributions, in order to prioritize descriptors for screening calculations. The results show that descriptor correlation effects, violating the assumption of feature independence, can lead to notable reduction of compound recall in Bayesian screening. Controlling descriptor correlation effects play a much more significant role for achieving high recall rates than approximating descriptor distributions by Gaussians. Furthermore, Kullback-Leibler divergence analysis is shown to systematically identify descriptors that are the most relevant for the outcome of Bayesian screening calculations.

INTRODUCTION

The development of methods to navigate high-dimensional descriptor spaces and search for active compounds has become a topical area of research in chemoinformatics.[1] Methods designed to analyze structure−activity relationships in high-dimensional space representations include support vector machines,[2,3] mapping algorithms,[4] or distance functions.[5] For example, an approach termed Distance in Activity-Centered Chemical Space (DACCS) has been introduced that centers high-dimensional space representations on subspaces defined by specific compound activity classes and calculates Euclidean distances in scaled chemical space between active and database compounds as a measure of molecular similarity.[5] Following this approach, increasing distance from the center of a subspace defined by a class of active compounds indicates decreasing similarity between active and database molecules. The DACCS method has then been transformed into a probabilistic measure on the basis of Bayesian principles.[6,7] The resulting BDACCS function expresses distance relationships in chemical space as a likelihood of activity;[7] the larger the distance between an active subspace and database compounds, the lower the probability that these compounds are active. BDACCS calculations involve the derivation of probability distributions of descriptor values of active and inactive compounds, and the divergence between these feature distributions can be calculated using the Kullback-Leibler (KL) function.[8] The logarithm of the KL-divergence of probability distributions between many different classes of active compounds and random database molecules has been found to correlate linearly with compound recovery rates of BDACCS screening calculations.[9] This linear model established by the KL-BDACCS formalism has been successfully applied to predict compound recall rates for Bayesian screening trials when compounds with diverse activities were used as reference molecules.[9]

Here we combine Kullback-Leibler analysis and the BDACCS method in a different way, namely by determining the KL-divergence for individual descriptors and estimating their relative importance for the outcome of search calculations. This makes it possible to identify small subsets of most relevant descriptors. With these different modifications, BDACCS calculations are found to produce significant recovery rates using small feature sets. Furthermore, we have set out to investigate the role of two fundamental approximations underlying the Bayesian screening process, i.e., the assumptions that individual descriptors are independent of each other and that descriptor value distributions follow a normal distribution. These approximations might compromise the accuracy of BDACCS screening. The role of the first approximation has been studied by generating sets of weakly correlated descriptors and comparing calculations in the presence and absence of significant correlation effects. The second approximation has been analyzed by introducing descriptor discretization schemes that alleviate the need to assume the presence of normal value distributions. The theoretical foundations of our analysis and the results are reported herein.

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

10.1021/ci700333t CCC: $40.75 © 2008 American Chemical Society
Published on Web 01/30/2008

METHODOLOGY

In our previous work[7,9] we have developed a statistical framework on the basis of Bayesian principles that makes it possible to assess the importance and the discriminatory power of molecular descriptors for a given compound activity class relative to a background database. This is accomplished by deriving a measure for the divergence of distributions of descriptor values for active compounds and a background database. These calculations are based on two principal assumptions: (1) descriptor distributions are independent of each other and (2) these distributions follow a normal distribution. For many practical applications, both assumptions represent approximations. Nevertheless, a distance measure based on these assumptions proved to be capable of retrieving active compounds at rates comparable to or higher than other descriptor-based methods.[7] Furthermore, considering the KL-divergence of descriptor distributions of active and database compounds, it has been possible to accurately estimate recovery rates of active compounds in Bayesian screening calculations.[9] This concept could be successfully extended to predict compound recovery rates for similarity searching using fingerprints that encode binary descriptor distributions.[10]

In this study, we modify and further evaluate this approach by considering the KL-divergence of individual descriptors for an activity class, thus producing a ranking of descriptors according to decreasing ability to discriminate between active and database compounds. Given this ranking scheme, one can assess how many descriptors are required to achieve high recovery rates in database searching.

The theoretical foundation of BDACCS is provided by the idea that descriptor values of compounds having similar activity might display a characteristic distribution that is different from random database compounds. Thus, if probability distributions are known for an activity class ($p(x|A)$) and a background database ($p(x|B)$), then the likelihood ratio can be expressed according to Bayes theorem[6] as

$$\frac{L(A \mid x)}{L(B \mid x)} = \alpha \frac{p(x \mid A)}{p(x \mid B)}$$

for a constant α where $x = (x_i)_{i = 1 \ldots n}$ represents the values of n descriptors for a compound x. This ratio determines the likelihood of a compound to be active. It is a relative measure of activity, and, therefore, it is not necessary to further consider the proportionality factor α. Comparing likelihood ratios of different compounds generates a ranking of these compounds according to the decreasing probability of activity. Taking the logarithm of the likelihood ratio produces a log-odds formulation

$$R(x) = -\log \frac{L(A \mid x)}{L(B \mid x)} = \log p(x \mid B) - \log p(x \mid A)$$

The expected value

$$E[-R(x)| A] = \int p(x \mid A) \log \frac{p(x \mid A)}{p(x \mid B)} \mathrm{d}x$$

is the KL-divergence

$$D\,[p(x \mid A) \,||\, p(x \mid B)]$$

of the distributions that reflects the ability of the distance functions to recall active compounds. Making the principal assumption of descriptor independence, the probability (or "pseudodistance") measure $R(x)$ is additive:

$$R(x) = \sum_{i=1}^{n} R(x_i) \text{ and } E[-R(x) \mid A] = \sum_{i=1}^{n} E[-R(x_i) \mid A]$$

Given this formula, the descriptors $i = 1 \ldots n$ can be ranked according to decreasing KL-divergence. Descriptors with large KL-divergence have high information content with respect to the activity class A and the population database B. If descriptors are sorted according to their KL-divergence, a set of "distance functions" can be defined:

$$R_m(x) = \sum_{i=1}^{m} R(x_i),\ m = 1 \ldots n$$

Using these functions for database searching we might generally expect increasing recall rates when adding more descriptors. However, descriptors having small KL-divergence do not significantly contribute to compound recall but are expected to mainly add "noise" to the scoring function.

Thus far, we do not make explicit assumptions about the underlying probability distributions, except that descriptors are thought to be independent. From a theoretical point of view, this independence assumption is an approximation inherent in the scoring functions $R_m(x)$. In order to exclude the possibility that strongly correlated descriptors might bias the calculations such descriptors should be eliminated. For this purpose, a method for selecting subsets of only weakly correlated descriptors can be applied termed unsupervised forward selection (UFS).[11] The UFS method iteratively selects descriptors that are the least correlated to already chosen descriptors until a predefined amount of the total variance of the descriptor distributions is accounted for by the resulting descriptor subset. For efficiency reasons, the UFS algorithm was reimplemented to exclusively operate on the cross-correlation descriptor matrix.

In our previous investigations, we have also made the principal assumption that the value distributions of all descriptors are normally distributed, which results in the following scoring function

$$R_m(x) = \sum_{i=1}^{m} \left(\log \frac{\sigma_i}{\tau_i} + \frac{(x_i - \mu_i)^2}{2\sigma_i^2} - \frac{(x_i - \nu_i)^2}{2\tau_i^2} \right)$$

where μ_i, σ_i are the estimated mean and the standard deviation of descriptor i for activity class A, and ν_i, τ_i are the mean and the standard deviation, respectively, for the population database B. It should be noted that this scoring function represents a modified normalized Euclidean distance function. Then the KL-divergence for these assumed Gaussians is

$$D_m[p(x \mid A) \,||\, p(x \mid B)] = \sum_{i=1}^{m} \left(\log \frac{\tau_i}{\sigma_i} + \frac{\sigma_i^2 - \tau_i^2 + (\mu_i - \nu_i)^2}{2\tau_i^2} \right)$$

We have been able to establish a relationship between these divergence measures and recall rates for Bayesian compound screening[9] and show that KL-divergence predicts the ability of descriptors to distinguish between active and inactive compounds with reasonable to high accuracy.

We now also evaluate an alternative to the assumption of normal distributions by discretizing descriptor values using a binning scheme and estimating the probability of occurrence for each bin for active and database compounds from the relative frequencies. As previously noted by Hsu et al.,[12] for continuous data, naïve discretization methods like equal-width binning generally performed well in the context of a Bayesian classification scheme when compared to more complex approaches.

For discretization, the probabilities

$$p(x_i \in Bin_i(k) \mid A) = p_i^A(k) \text{ and } p(x_i \in Bin_i(k) \mid B) = p_i^B(k)$$

are estimated from the relative frequencies with which values of descriptor i fall within bin k. A Laplacian correction is used to smooth the relative frequencies and avoid cases where a probability becomes 0. As a binning scheme, we apply simple equidistant binning into 5 bins ranging from the 0.001 quantile to 0.999 quantile range of each descriptor value distribution in the background database in order to eliminate the influence of outlier values on the bin ranges. Then the incremental scoring function $R_m(x)$ becomes

$$R_m(x) = \sum_{i=1}^{m} R(x_i) = \sum_{i=1}^{m} (\log p_i^B(x_i) - \log p_i^A(x_i)),\ m = 1 \ldots n$$

and the KL-divergence is written as

$$D_m[p(x \mid A) \,||\, p(x \mid B)] = \sum_{i=1}^{m} \sum_{k=1}^{5} p_i^A(k) \log \frac{p_i^A(k)}{p_i^B(k)}$$

Thus, with the modifications described above, we have been able to evaluate the fundamental assumptions of descriptor independence and normal value distributions. Furthermore, we now incrementally add descriptor contributions to KL-BDACCS analysis and evaluate the consequences on compound recall in screening calculations.

BENCHMARK CALCULATIONS

Screening calculations using BDACCS were performed on 37 previously reported[9] compound activity classes (summarized in Supporting Information Table 1). From each class, 100 sets of 20 reference molecules were randomly selected, and the remaining compounds were added to a subset of ZINC[13] that was used as the background database, containing approximately 3.7 million compounds after applying in-house druglikeness filters to the 5.6 million compounds of version 7 of ZINC. For each series of calculations, 100 individual screening trials were carried out per class so that the results were statistically sound and not biased by reference set composition or other chance effects. Average results are reported. As descriptors, a pool of 142 molecular property descriptors available in the Molecular Operating Environment (MOE)[14] was used, as described previously.[7]

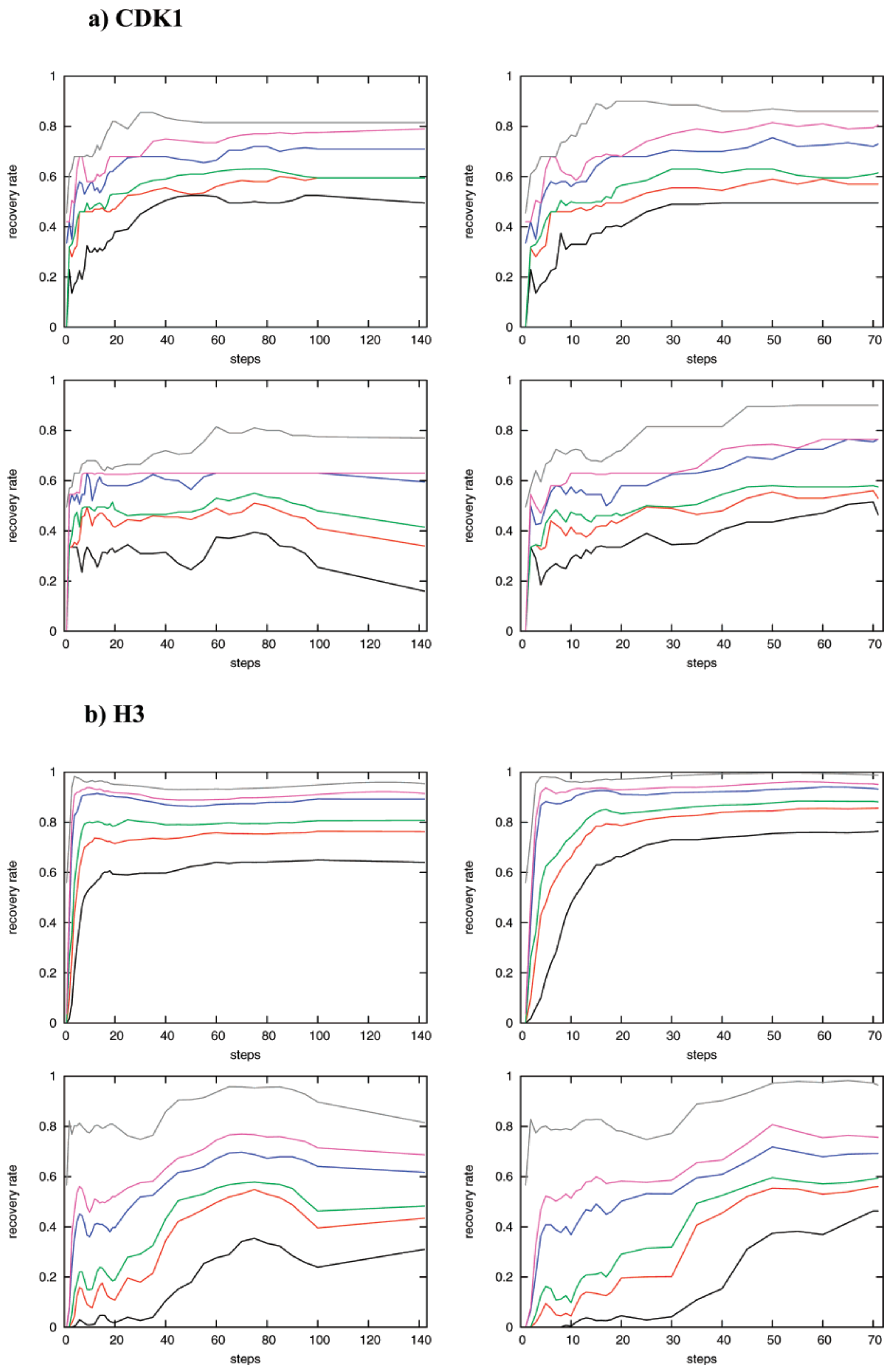
The assumption of descriptor independence was assessed using descriptor sets having different degrees of intraset correlation (few, if any molecular properties descriptors are truly uncorrelated). In one series of calculations, all 142 descriptors were used, including a number of strongly correlated ones. In another series, the UFS method was applied to select the 71 least correlated descriptors, i.e., 50% of the pool accounting for 98% of the total variation in the background database. The descriptors in the 142- and 71-set were ranked according to KL-divergence between active and database compounds, and BDACCS calculations on all 37 activity classes were carried out with incrementally extended descriptor sets: the top ranked 20 descriptors were added one at a time, and descriptors ranked from 21 to 100 (or 71) were added in sets of five. Finally, the remaining 42 descriptors were included. These series of calculations made it possible to evaluate the screening performance of descriptors selected on the basis of KL-divergence analysis and the influence of descriptor correlation effects in parallel.

The assumption of normal descriptor value distributions was evaluated by using discretized descriptors. For discretization, the relatively small number of reference compounds available for screening calculations presents a statistical complication. Therefore, in order to estimate relative frequencies of descriptor values well, it was necessary to test different numbers of bins for discretizing descriptor value ranges. For 20 reference compounds, discretization of descriptor value ranges through equidistant binning over five intervals gave the overall best compound recall. Again, two series of calculations were carried out using discretized descriptors, one with all 142 descriptors and the other using the UFS-selected 71 descriptor set.

For the analysis of compound recall, recovery rates were calculated and graphically analyzed using cumulative recall curves for the top-ranked 100, 500, 1000, 5000, 10 000, and 50 000 database compounds; the latter two subsets correspond to ~0.27% and ~1.37% of the source database, respectively.

RESULTS AND DISCUSSION

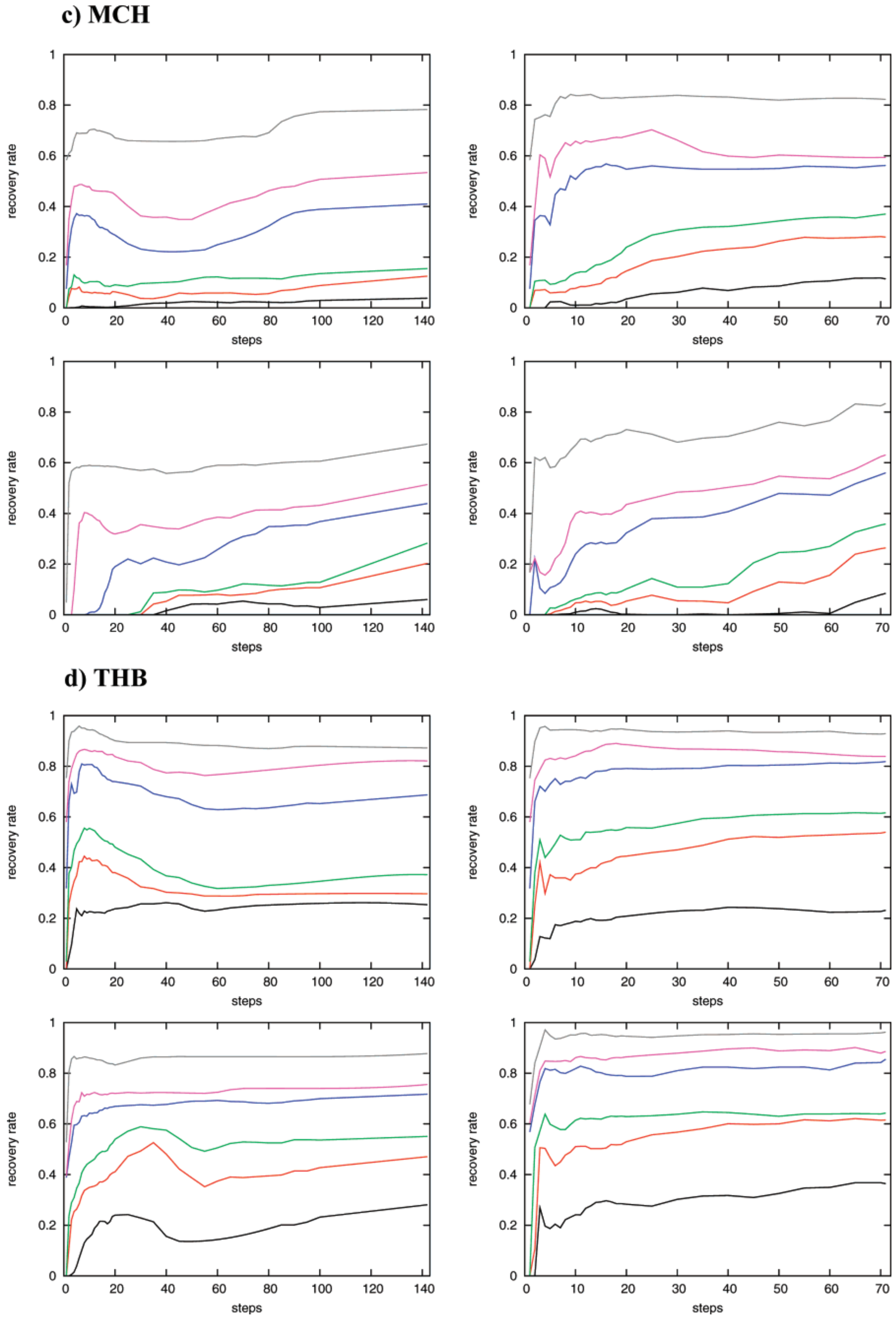
We have generated different versions of the BDACCS method in order to evaluate the influence of descriptor correlation and discretization on Bayesian screening. Descriptor correlation and discretization analyses address the assumptions of feature independence and normal value distributions, respectively, which underlie the BDACCS approach. Furthermore, we have carried out KL-divergence analysis for individual descriptors in order to prioritize them for screening and study the performance of descriptor sets of increasing size. Systematic screening calculations were performed on a total of 37 different compound activity classes. Figure 1 shows cumulative compound recall under variation of descriptor sets for representative examples of classes that displayed different behavior in screening trials. Figure 2 compares the performance of the different BDACCS versions for representative activity classes. Table 1 lists descriptors that were overall preferred under varying BDACCS calculation conditions, i.e., most significant on the basis of KL-divergence analysis. Finally, in Table 2, we report recovery rates of BDACCS calculations on all 37 activity classes. The Supporting Information includes recall curves and performance comparisons for all remaining activity classes and the minimum and maximum recall rates for the values reported in Table 2.



a) CDK1

b) H3

recovery rate
steps
0
0.2
0.4
0.6
0.8
1
0
20
40
60
80
100
120
140
0
10
20
30
40
50
60
70



c) MCH

d) THB

Figure 1. Cumulative compound recall under descriptor variation. Average recovery rates are plotted against the number of descriptors used in BDACCS screening trials. Descriptors are added in decreasing order of significance estimated by KL-divergence. "Steps" reports the stepwise increasing number of descriptors per calculation. Graphs are shown for the top 100 (black), 500 (red), 1000 (green), 5000 (blue), 10 000 (magenta), and 50 000 (gray) compounds for four different versions of the BDACCS method: continuous descriptor values for 142 descriptors (top left); continuous descriptor values for 71 descriptors selected by the UFS method (top right); 142 discretized descriptors (bottom left); and 71 discretized descriptors selected by UFS (bottom right). a) Results for cyclin-dependent kinase 1 inhibitors (CDK1), b) H3 antagonists, c) melanin-concentrating hormone (MCH), and d) thrombin inhibitors (THB).

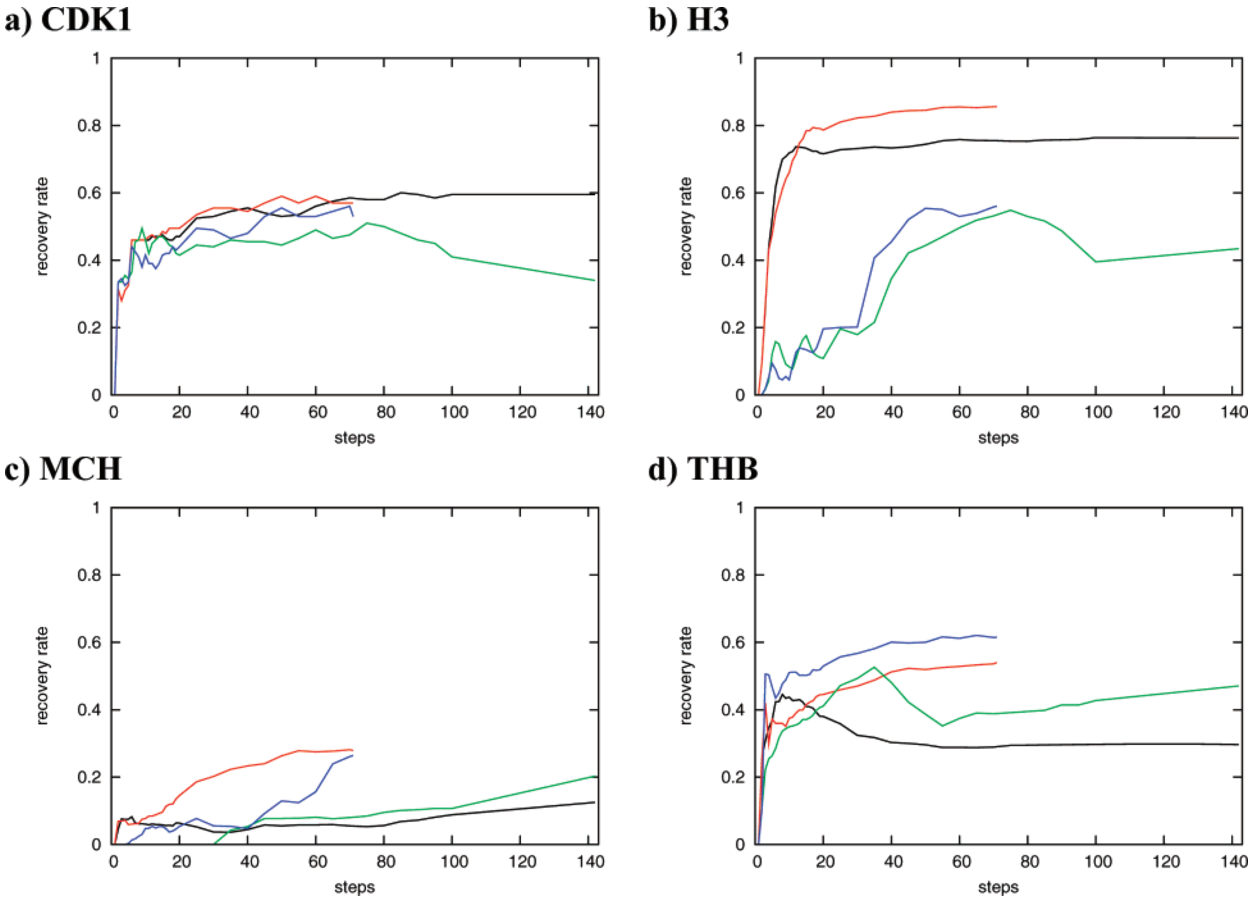


Figure 2. Direct comparison of different BDACCS versions. Average recovery rates of the top 500 compounds are plotted against the number of descriptors for the four activity classes as in Figure 1. Recovery rates achieved by the four different versions of BDACCS are plotted in a single diagram: continuous descriptor values for 142 descriptors (black), continuous descriptor values for 71 descriptors selected by UFS (red), 142 discretized descriptors (green), and 71 discretized descriptors selected by UFS (blue).

Figure 1 and Supporting Information Figure 1 show in part significant differences in recall of active compounds for different activity classes. The compound class-dependence of virtual screening calculations is a well-known phenomenon[1] and typically observed independent of the methods that are used.[1] Specific molecular features that might determine such effects are currently not known. In order to provide a meaningful basis for a thorough assessment of the methodologies reported herein, we have therefore investigated a large number of different compound classes. Our 37 activity classes cover a wide range of different chemotypes with varying intraset molecular diversity, which ensures that the overall results are not strongly influenced by individual compound classes.

In addition to expected differences in recovery rates, the activity classes responded in part very differently to varying screening conditions and produced different phenotypes, as shown in Figures 1 and 2. In some cases, a significant enrichment of active molecules was observed in small selection sets (Figure 1a,b), in contrast to others where larger selection sets were required (Figure 1c,d). Independent of the magnitude of recovery rates, the significance of descriptors played a critically important role for compound recall in the majority of cases. Figure 1 shows that top recovery rates were often observed for BDACCS calculations using only approximately 10 to 30 descriptors and that recall curves already reached a plateau after relatively few descriptors were added. In some cases, these effects were dramatic, depending on the calculation conditions (e.g., Figure 1b,d). These findings revealed that relatively small subsets of descriptors frequently produced top recovery rates. For BDACCS calculations, KL-divergence was a reliable indicator of descriptor significance. Table 1 lists descriptors that were most important for screening calculations using the four BDACCS versions. Among preferred descriptors, BCUT-type descriptors[15] and descriptors combining molecular surface and property information[16] were prevalent. These types of descriptors have in common that they are designed to be orthogonal and information-rich. However, less complex descriptors were also found among the most significant ones including, for example, connectivity indices or simple atom counts. Table 1 also shows that many top-ranked descriptors in the 142-set were not selected by the UFS method, due to significant correlation with others.

Figure 2 further illustrates that BDACCS calculation conditions had different effects on compound recall. In some cases, differences in descriptor correlation and continuous or discretized descriptor values had only a little effect on recovery rates (Figure 2a), but in others, rates differed in significant ways. For example, for class H3 (Figure 2b), BDACCS calculations with continuous descriptor values were strongly preferred, irrespective of differences in correlation. However, these trends were reversed in other cases. For example, class THB (Figure 2d) displayed a clear preference for descriptor sets with limited correlation,

Table 1. Most Significant Descriptors According to the KL-Divergence[a]

			no. of activity classes for which descriptor is ranked in the top x list						no. of activity classes for which descriptor is ranked in the top x list		
rank	descriptor	descriptor type	top 5	top 10	top 20	rank	descriptor	descriptor type	top 5	top 10	top 20
					a) 142 Continuous Descriptors						
1	weinerPath	connectivity index	10	12	14	11	SlogP_VSA1	CSA	4	6	8
2	chiral	atom count	7	10	13	12	PEOE_VSA+6	CSA	4	6	7
3	SlogP_VSA0	CSA	6	7	13	13	GCUT_SLOGP_3	BCUT-type	4	6	6
4	vsa_don	CSA	6	7	9	14	a_don	atom count	4	6	6
5	a_nN	atom count	5	7	10	15	PEOE_VSA-6	CSA	4	5	8
6	PEOE_PC-*	physical property	5	6	10	16	SMR_VSA0*	CSA	3	8	12
7	BCUT_SLOGP_0	BCUT-type	5	6	8	17	a_hyd*	atom count	3	8	10
8	SMR_VSA4	CSA	5	5	7	18	vsa_acc	CSA	3	7	11
9	GCUT_SLOGP_0	BCUT-type	4	6	12	19	KierA1*	connectivity index	3	7	9
10	GCUT_SMR_0	BCUT-type	4	6	9	20	SMR_VSA1	CSA	3	5	9
					b) 71 UFS-Selected Continuous Descriptors						
1	weinerPath	connectivity index	12	17	19	11	a_nN	atom count	6	11	18
2	chiral	atom count	10	13	18	12	PEOE_VSA+6	CSA	6	9	14
3	SlogP_VSA0	CSA	7	14	19	13	PEOE_VSA-6	CSA	6	8	13
4	GCUT_SLOGP_3	BCUT-type	7	11	18	14	SMR_VSA4	CSA	6	7	11
5	GCUT_SLOGP_0	BCUT-type	7	10	17	15	SlogP_VSA8	CSA	5	12	19
6	vsa_don	CSA	7	9	15	16	KierA3	connectivity index	5	11	20
7	GCUT_SMR_0	BCUT-type	7	9	11	17	opr_nrot	bond count	5	8	21
8	BCUT_SLOGP_0	BCUT-type	7	8	12	18	PEOE_VSA+1	CSA	5	7	10
9	a_don	atom count	7	7	10	19	SlogP_VSA1	CSA	4	12	17
10	vsa_acc	CSA	6	14	22	20	SMR_VSA1	CSA	4	10	12
					c) 142 Discretized Descriptors.						
1	PEOE_VSA_FHYD*	CSA	7	9	11	11	vsa_don	CSA	4	5	7
2	a_hyd*	atom count	6	8	12	12	KierA1*	connectivity index	4	4	7
3	TPSA*	physical property	6	7	9	13	SMR_VSA5*	CSA	3	6	13
4	opr_brigid*	bond count	5	5	9	14	SMR_VSA0*	CSA	3	6	8
5	BCUT_PEOE_2	BCUT-type	4	12	13	15	Kier1*	connectivity index	3	6	8
6	SlogP_VSA0	CSA	4	7	9	16	rings*	atom count	3	6	7
7	chi0_C*	connectivity index	4	6	9	17	chi1v*	connectivity index	3	5	10
8	opr_nring*	atom count	4	6	7	18	a_nC*	atom count	3	5	9
9	GCUT_SMR_3*	BCUT-type	4	5	8	19	GCUT_SLOGP_3	BCUT-type	3	5	8
10	a_don	atom count	4	5	8	20	GCUT_PEOE_3*	BCUT-type	3	5	8
					d) 71 UFS-Selected Discretized Descriptors						
1	weinerPath	connectivity index	11	11	14	11	vsa_don	CSA	6	6	13
2	BCUT_PEOE_2	BCUT-type	10	18	20	12	BCUT_SMR_3	BCUT-type	6	6	10
3	SlogP_VSA0	CSA	10	13	16	13	SMR_VSA1	CSA	5	9	10
4	b_ar	bond count	9	13	14	14	GCUT_PEOE_2	BCUT-type	5	7	13
5	radius	connectivity index	8	13	15	15	PEOE_VSA-6	CSA	5	7	10
6	logP(o/w)	physical property	8	12	16	16	SMR_VSA4	CSA	5	6	10
7	GCUT_SLOGP_3	BCUT-type	7	10	13	17	PEOE_VSA-3	CSA	4	12	28
8	opr_nrot	bond count	6	10	20	18	PEOE_VSA-1	CSA	4	11	13
9	vsa_acc	CSA	6	10	14	19	a_nN	atom count	4	9	12
10	a_don	atom count	6	10	11	20	SlogP_VSA2	CSA	4	8	8

[a] For each BDACCS version, descriptors are ranked for each activity class according to their KL-divergence. Descriptors that occur most frequently within the top 5, 10, and 20 are shown for each version. Descriptors marked with an asterisk were not selected by the UFS method and were thus not part of the 71 descriptor selection set. Descriptors are classified as physical property, atom/bond count, connectivity index, BCUT-type,[15] or complex surface area-type[16] (CSA) descriptors. Abbreviations are used according to MOE[14] with which descriptor values were calculated.

regardless of whether their value ranges were continuous or binned.

For in silico screening, the estimation of differences in feature distributions is principally complicated by the fact that the number of available reference molecules is usually small. Moreover, for small reference sets, the nature of their descriptor value distributions is difficult to predict. In principle, descriptor discretization is capable of accounting for departures from Gaussian distributions. However, for small reference sets, only a small number of bins can be used to represent value frequencies in a meaningful way. Furthermore, independent of reference set size, the assumption of feature independence can only be addressed by using sets of more or less correlated descriptors because truly uncorrelated descriptors can hardly be identified. However, our systematic benchmark calculations revealed some clear trends, as shown in Table 2 that lists recovery rates for all compound classes and BDACCS versions using selection sets of 500 database compounds as a reference point. In only one of 37 test cases, BDACCS calculations failed to recover any active molecules among the top 500 database compounds. Importantly, for 25 of 37 classes, the best recovery rates were observed for UFS-selected descriptors with continuous value ranges. In nine other cases, however, UFS-selected descriptors performed best when their value ranges were discretized. Only two classes displayed best recovery rates for 142 continuous descriptors. When these 142 descriptors were discretized, top hit rates were never observed. Taken together, these findings showed that descriptor correlation effects generally had a strong influence

Table 2. Recovery Rates[a]

	continuous		discretized	
activity classes	all *mean % (SD)*	UFS *mean % (SD)*	all *mean % (SD)*	UFS *mean % (SD)*
5HT	38.22 (7.66)	**45.80** (9.73)	18.94 (8.17)	26.61 (6.34)
AA2	1.87 (3.68)	6.13 (5.42)	5.07 (5.20)	**8.87** (4.93)
ANA	28.80 (10.06)	**66.20** (7.15)	25.80 (7.70)	44.56 (6.85)
ARI	10.00 (13.30)	**14.50** (17.47)	10.00 (12.81)	8.25 (11.81)
ARO	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
BEN	38.28 (9.39)	**57.62** (11.03)	23.36 (6.48)	30.49 (8.75)
CA	22.30 (9.59)	31.38 (7.53)	18.70 (6.38)	**34.29** (5.92)
CAL	1.88 (5.44)	**13.75** (11.31)	0.00 (0.00)	11.25 (8.97)
CDK1	**59.50** (33.86)	57.00 (34.10)	34.00 (32.47)	53.00 (33.95)
CDK2	52.50 (23.70)	**60.00** (23.84)	16.75 (15.51)	58.50 (24.67)
CHO	3.60 (7.46)	**9.30** (11.30)	0.00 (0.00)	0.00 (0.00)
COX	85.91 (8.41)	**87.73** (7.79)	84.55 (8.22)	84.09 (10.20)
DD1	54.40 (12.17)	59.70 (11.67)	59.40 (10.81)	**65.90** (11.20)
DIR	24.90 (11.50)	**32.40** (12.32)	0.00 (0.00)	11.60 (8.49)
EDN	23.50 (10.88)	**41.67** (12.19)	29.17 (11.39)	34.67 (11.53)
ESU	3.07 (5.05)	29.73 (10.05)	3.47 (4.88)	**33.93** (10.85)
GLY	32.71 (13.96)	**61.29** (13.48)	17.00 (8.90)	51.00 (11.75)
GRH	43.18 (3.18)	63.40 (9.01)	42.15 (5.73)	**63.74** (5.88)
H1D	**15.88** (8.21)	15.38 (7.87)	6.06 (5.58)	5.38 (5.33)
H3	76.28 (7.55)	**85.62** (6.85)	43.44 (9.35)	56.06 (5.97)
HIV	81.93 (6.92)	91.14 (4.17)	85.75 (6.76)	**92.32** (3.71)
INO	35.93 (11.29)	**48.20** (13.26)	0.20 (1.14)	10.87 (11.20)
JNK	3.00 (3.14)	**22.62** (7.47)	6.19 (4.74)	16.31 (8.04)
KAP	12.20 (13.90)	**21.80** (16.84)	4.60 (8.92)	9.20 (12.20)
KRA	15.50 (27.24)	**22.00** (26.89)	0.00 (0.00)	0.00 (0.00)
LAC	26.00 (11.52)	**31.22** (12.60)	0.00 (0.00)	17.89 (11.04)
LDL	10.20 (8.53)	22.70 (11.09)	18.60 (10.54)	**25.00** (11.85)
LIP	12.57 (7.61)	**20.43** (10.94)	0.00 (0.00)	0.00 (0.00)
MCH	12.50 (8.69)	**27.90** (11.49)	20.30 (9.79)	26.40 (10.40)
MEL	28.00 (19.49)	**29.20** (19.98)	10.60 (14.34)	0.60 (3.43)
SQS	45.27 (7.36)	**52.18** (7.44)	48.18 (7.37)	48.00 (7.52)
THI	3.29 (4.47)	**5.07** (5.86)	0.00 (0.00)	3.21 (3.98)
THR	16.46 (11.52)	42.00 (11.12)	15.62 (9.43)	**42.31** (11.39)
TK	89.00 (9.68)	**96.60** (4.39)	62.73 (16.95)	84.07 (9.66)
THB	29.67 (8.50)	54.00 (10.62)	47.07 (11.00)	**61.53** (9.37)
VEG	28.88 (10.87)	**32.75** (9.53)	20.00 (7.38)	6.62 (6.64)
XAN	42.40 (10.18)	**51.93** (9.49)	26.13 (9.60)	26.73 (8.50)

[a] For 37 activity classes (abbreviated according to Supporting Information Table 1), average compound recovery rates are reported for the top 500 database compounds over 100 trials for the four different versions of the BDACCS method. Best recall rates for each class are shown in bold.

on the quality of BDACCS calculations. Also, differences in descriptor correlation were more important than differences between continuous and discretized value distributions. However, differences between continuous and discretized descriptors also played a significant role in a number of cases, as illustrated by many of the 25 activity classes where continuous UFS-selected descriptors produced best recovery rates. In nine of these cases, descriptor discretization reduced recovery rates by more than 10%, and, in three cases, calculations with discretized descriptors failed to recover active compounds. By contrast, for nine other classes, the screening performance of UFS-selected descriptors further increased when their value ranges were discretized. These increases were due to better representation of irregular distributions of reference set descriptors. Taken together, the screening results suggest that continuous value distributions overall provided a better model for BDACCS screening than discretized distributions, although discretization led to notable increases in recovery rates in a number of cases. Thus, underlying approximations influenced the outcome of BDACCS calculations in different ways.

Table 3. Statistical Analysis of Screening Performance[a]

	continuous all	continuous UFS	discretized all	discretized UFS
continuous all	−	0	22	14
continuous UFS	29	−	33	20
discretized all	9	0	−	2
discretized UFS	16	6	26	−

[a] For 37 activity classes recall rates were compared in statistical tests between different methods. The figures show the number of classes where the method designated on the left performed better than the method designated on the top at a significance level of $\alpha = 0.01$.

For each of the 37 activity classes we performed statistical tests to assess the relative performance of BDACCS with or without UFS descriptor selection as well as using either continuous or discretized values. For the 100 trials performed per test case recall rates for the top 500 compounds between two different methods were compared using a nonparametric sign test. Following a conservative approach, trials with identical recall rates were evenly split into positive and negative signs, and the significance level was set at $\alpha = 0.01$. The results for all 37 activity classes are summarized in Table 3. As can be seen, selecting descriptors with the UFS method generally yields a significant improvement over nonselection. For continuous descriptors, 29 classes performed significantly better using descriptor selection. Moreover, there was not a single case where BDACCS performed better without selection. Discretized descriptors show a similar behavior: for 26 activity classes, better results were achieved using descriptor selection, as opposed to two cases for which discretized descriptors without selection were preferred. Overall we found that in 16 cases the combination of continuous descriptors and UFS selection performed significantly better than any other method, whereas the combination of discrete descriptors and UFS performed best in six cases.

CONCLUDING REMARKS

The major objectives of our study have been to evaluate Kullback-Leibler divergence analysis in order to prioritize and select descriptors for BDACCS calculations and to explore the role of fundamental assumptions underlying the Bayesian screening approach: feature independence and the presence of normal value distributions. Therefore, four variants of the BDACCS method were designed and compared in systematic screening trials on 37 compound activity classes. KL-divergence was a reliable indicator of descriptor significance for BDACCS screening and small to moderately sized descriptor sets often performed best. By contrast, the addition of less significant descriptors frequently reduced recovery rates, which could be attributed to descriptor correlation effects. Thus, limited numbers of descriptors with high KL-significance and little correlation are preferred for BDACCS screening. The assumption of feature or descriptor independence is violated by descriptor correlation. Our analysis has shown that these effects can reduce screening performance in a significant way. Furthermore, systematic screening trials indicate that the assumption of normal value distributions represents a less significant approximation than feature independence. Taken together, our findings emphasize the need for descriptor correlation and significance analysis to optimize Bayesian screening performance. For descriptor selection, KL-divergence analysis should also be a useful method for applications other than BDACCS calculations. The analysis can be generally applied to estimate the significance of descriptors to, for example, account for differences between compound classes, data sets, or libraries or to select suitable descriptors for other applications.

Supporting Information Available: Figures 1 and 2 and Tables 1 and 2. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Eckert, H.; Bajorath, J. Molecular Similarity Analysis in Virtual Screening: Foundations, Limitations and Novel Approaches. *Drug Discovery Today* **2007**, *12*, 225−233.

(2) 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.

(3) Jorissen, R. N.; Gilson, M. K. Virtual Screening of Molecular Databases Using a Support Vector Machine. *J. Chem. Inf. Model.* **2005**, *45*, 549−561.

(4) Eckert, H.; Vogt, I.; Bajorath, J. Mapping Algorithms for Molecular Similarity Analysis and Ligand-based Virtual Screening: Design of DynaMAD and Comparison with MAD and DMC. *J. Chem. Inf. Model.* **2006**, *46*, 1623−1634.

(5) 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.

(6) Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*, 2nd ed.; Wiley-Interscience: New York, 2000; pp 20−83.

(7) Vogt, M.; Godden, J. W.; Bajorath, J. Bayesian Interpretation of a Distance Function for Navigating High-Dimensional Descriptor Spaces. *J. Chem. Inf. Model.* **2007**, *47*, 39−46.

(8) Kullback, S. *Information Theory and Statistics*; Dover Publications: Mineola, MN, 1997; pp 1−11.

(9) Vogt, M.; Bajorath, J. Introduction of an Information-Theoretic Method to Predict Recovery Rates of Active Compounds for Bayesian in Silico Screening: Theory and Screening Trials. *J. Chem. Inf. Model.* **2007**, *47*, 337−341.

(10) Vogt, M.; Bajorath, J. Introduction of a Generally Applicable Method to Estimate Retrieval of Active Molecules for Similarity Searching Using Fingerprints. *ChemMedChem* **2007**, *2*, 1311−1320.

(11) Whitley, D. C.; Ford, M. G.; Livingstone, D. J. Unsupervised Forward Selection: A Method for Eliminating Redundant Variables. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1160−1168.

(12) Hsu, C. N.; Huang, H. J.; Wong, T. T. Why discretization works for naïve Bayesian classifiers. Proceedings of the 17th International Conference on Machine Learning (ICML-2000) Stanford, CA, 2000; Langley, P., Ed.; Morgan Kaufmann: San Francisco, CA, pp 399−406.

(13) 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.

(14) *Molecular Operating Environment (MOE), Vers. 2005.06*; Chemical Computing Group Inc.: Suite 910 − 1010 Sherbrooke St. W, Montreal, Quebec, Canada H3A 2R7. http://www.chemcomp.com (accessed Nov 1, 2006).

(15) Pearlman, R. S.; Smith, K. M. Novel Software Tools for Chemical Diversity. *Perspect. Drug Discovery Des.* **1998**, *9*, 339−353.

(16) Labute, P. Derivation and Applications of Molecular Descriptors Based on Approximate Surface Area. *Methods Mol. Biol.* **2004**, *275*, 261−278.

CI700333T