Development of a Fingerprint Reduction Approach for Bayesian Similarity Searching Based on Kullback−Leibler Divergence Analysis

Britta Nisius, Martin Vogt, and Jürgen Bajorath*

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

Received March 5, 2009

The contribution of individual fingerprint bit positions to similarity search performance is systematically evaluated. A method is introduced to determine bit significance on the basis of Kullback−Leibler divergence analysis of bit distributions in active and database compounds. Bit divergence analysis and Bayesian compound screening share a common methodological foundation. Hence, given the significance ranking of all individual bit positions comprising a fingerprint, subsets of bits are evaluated in the context of Bayesian screening, and minimal fingerprint representations are determined that meet or exceed the search performance of unmodified fingerprints. For fingerprints of different design evaluated on many compound activity classes, we consistently find that subsets of fingerprint bit positions are responsible for search performance. In part, these subsets are very small and contain in some cases only a few fingerprint bit positions. Structural or pharmacophore patterns captured by preferred bit positions can often be directly associated with characteristic features of active compounds. In some cases, reduced fingerprint representations clearly exceed the search performance of the original fingerprints. Thus, fingerprint reduction likely represents a promising approach for practical applications.

INTRODUCTION

Similarity searching using molecular fingerprints is a widely applied technique to mine databases for active compounds.[1−3] Although fingerprints generally share a bit string design to represent structural features and molecular properties, diverse types of fingerprints have been introduced that encode molecular information in very different ways, for example, as substructures, topological pathways, pharmacophore patterns, or numerical property descriptors.[2] Although fingerprints might include feature counts, molecular information is most commonly represented in a binary format via bit positions that account for the presence or absence of features. Furthermore, most fingerprints have a fixed length, but molecule- or compound class-specific fingerprints of variable format have also been introduced.[4,5] Regardless of fingerprint design, similarity searching generally relies on the use of complete fingerprint representations as descriptors and on the quantification of fingerprint overlap as a measure of molecular similarity.[1] Contributions of individual bit positions have also been considered in fingerprint searching, for example, through generation of consensus fingerprints,[6] bit scaling,[7] reverse fingerprinting,[8] or bit silencing.[9] These techniques typically emphasize individual bit positions on the basis of bit frequency analysis to tune search calculations toward specific compound classes but do not modify the fingerprint format. Only recently, attempts have been made to perform "bit surgery" on the basis of frequency analysis and reduce the size of fingerprints by eliminating positions that do not contribute to search performance or hinder the preferential recognition of active compounds.[10] For example, the difference between bit silencing and bit reduction is that silencing sets individual "1" bits in fingerprints to "0" and hence negates their influence on the calculation of similarity coefficient values, whereas bit reduction eliminates bit positions and the corresponding features from fingerprints. However, both bit silencing and bit reduction methods have indicated that a number of bits in different fingerprints negatively affect their search performance.[9,10] In this study, we follow up on this theme and develop a specific fingerprint reduction method that is based on Bayesian and information-theoretic principles in order to determine minimal fingerprint representations that maintain or exceed the search performance of unmodified fingerprints.

In a previous work, the development of a statistical framework based on Bayesian principles was established to assess the importance and discriminatory power of numerical molecular descriptors for a given compound activity class relative to a background database.[11] Furthermore, as a measure of the divergence of descriptor value distributions of active compounds and database molecules, the Kullback−Leibler (KL) divergence was adopted from information theory.[12] Considering the KL divergence of descriptor distributions of active and database compounds, it was possible to correctly estimate recovery rates of active compounds in Bayesian screening calculations[12] and introduce a descriptor ranking procedure based on KL divergences.[13] Furthermore, this concept was successfully extended to predict compound recovery rates for similarity searching using fingerprints[14] and rank different fingerprints according to their ability to successfully detect compounds belonging to a specific activity class in a similarity search.[15]

Here, we extend the KL divergence analysis approach by considering KL divergences of individual fingerprint bits for

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

10.1021/ci900087y CCC: $40.75
Published on Web 05/29/2009

a compound activity class, thus producing a bit ranking scheme according to decreasing ability to discriminate between active and database compounds. On the basis of this ranking, one can generate and evaluate fingerprints of reduced size to determine how many bits are required to achieve high recovery rates in similarity searching and, in addition, whether bit reduction can lead to an increase in the recovery of active compounds. We find that only subsets of fingerprint bits are responsible for screening performance, regardless of the type of fingerprint, and that reduced fingerprint representations might achieve higher recovery rates than the original fingerprints.

METHODOLOGY

Fingerprints $\vec{v} = (v_i)_{i=1,\ldots,k}$ consist of k bits that might either be set on or off. Therefore, fingerprint bits can be modeled as random binary variables following a Bernoulli distribution. The Bernoulli distribution is defined by a single parameter p describing the probability that the variable is set on. Accordingly, the probability that a bit is set off is given by $q = 1 - p$. The probabilities $p_i, i = 1, \ldots, k$ can be estimated from the relative frequencies with which an individual bit is set on within a class of active compounds (A) or the background database (B)

$$p_i^A = \frac{\#\{\vec{v} \in A | v_i = 1\}}{m} \text{ and } p_i^B = \frac{\#\{\vec{v} \in B | v_i = 1\}}{n} \tag{1}$$

with m denoting the number of active and n the number of database compounds. Estimating the probabilities based on relatively small training sets can result in inaccurate probability distributions. In order to avoid such effects and balance the distributions, an m-estimate correction[16] is applied that adds a single compound to the active training data reflecting the bit settings in the compound database. An equivalent correction can be applied to the inactive background database

$$\hat{p}_i^A = \frac{mp_i^A + p_i^B}{m+1}, \hat{p}_i^B = \frac{np_i^B + p_i^A}{n+1} \tag{2}$$

Assuming that different fingerprint bits are independent, the probability $P(\vec{v}|A)$ can be calculated by

$$P(\vec{v}|A) = \prod_{i=1}^{k} P(v_i|A) = \prod_{i=1}^{k} (\hat{p}_i^A)^{v_i} (\hat{q}_i^A)^{1-v_i} \tag{3}$$

and, in analogy, for $P(\vec{v}|B)$.

According to Bayes theorem the likelihood ratio can be expressed as

$$R(\vec{v}) = \frac{L(A|\vec{v})}{L(B|\vec{v})} \propto \frac{P(\vec{v}|A)}{P(\vec{v}|B)} = \prod_{i=1}^{k} \frac{(\hat{p}_i^A)^{v_i} (\hat{q}_i^A)^{1-v_i}}{(\hat{p}_i^B)^{v_i} (\hat{q}_i^B)^{1-v_i}} \tag{4}$$

Taking the logarithm and simplifying the expression results in the following fitness function

$$\log(R(\vec{v})) = \sum_{i=1}^{k} v_i \left(\log \frac{\hat{p}_i^A}{\hat{p}_i^B} - \log \frac{\hat{q}_i^A}{\hat{q}_i^B} \right) + const \tag{5}$$

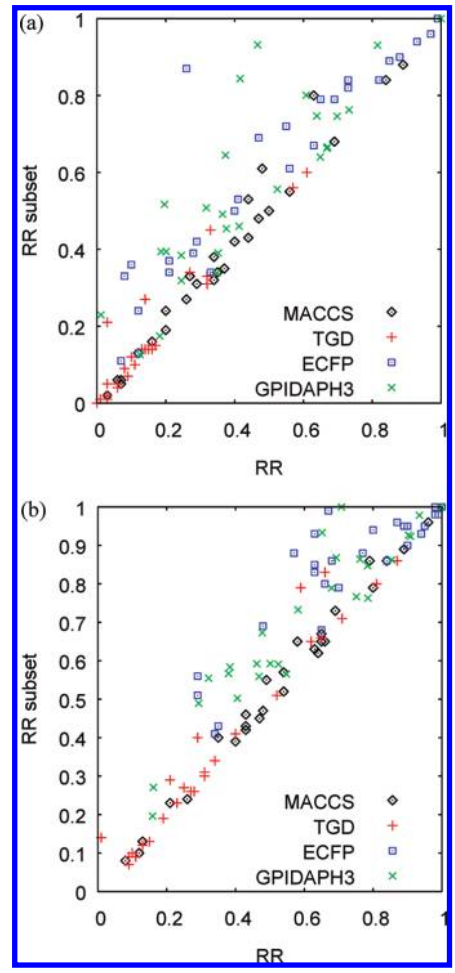

Figure 1. Comparison of recovery rates for full-length fingerprints and best bit subset. Results are shown for database selection sets of (a) 100 and (b) 1000 compounds. "RR" corresponds to recovery rates for full-length fingerprints, and "RR subset" corresponds to the best-performing reduced fingerprint.

that corresponds to a weighting scheme (R4) for fingerprints introduced by Ormerod et al.[17] and studied in detail by Hert et al.[18]

To evaluate the ability of an individual bit position to discriminate between active and database compounds, the probability distributions of this bit within active compounds and the database must be compared. Therefore, we calculate the KL divergence from information theory[19] to quantify the difference between these two probability distributions

$$D[p(\vec{v}|A)||p(\vec{v}|B)] = \sum_{\vec{b}=(0,\ldots,0)}^{(1,\ldots,1)} P(\vec{v} = \vec{b}|A) \log \frac{P(\vec{v} = \vec{b}|A)}{P(\vec{v} = \vec{b}|B)} \triangleq \sum_{i=1}^{k} \left(\hat{p}_i^A \log \frac{\hat{p}_i^A}{\hat{p}_i^B} + \hat{q}_i^A \log \frac{\hat{q}_i^A}{\hat{q}_i^B} \right) \tag{6}$$

This divergence is a measure for the discriminatory power of each bit position in similarity searching and can be utilized to rank fingerprint bits according to their relative importance.

The likelihood ratio in eq 5 on which the divergence analysis is based can be utilized for similarity searching. It assigns a log-odds score to a fingerprint $\vec{v} = (v_i)_{i=1,\ldots,k}$ of a given compound. In this context, the KL divergence in eq 6 corresponds to the expected score for an active compound. Thus, bit positions making large contributions to global

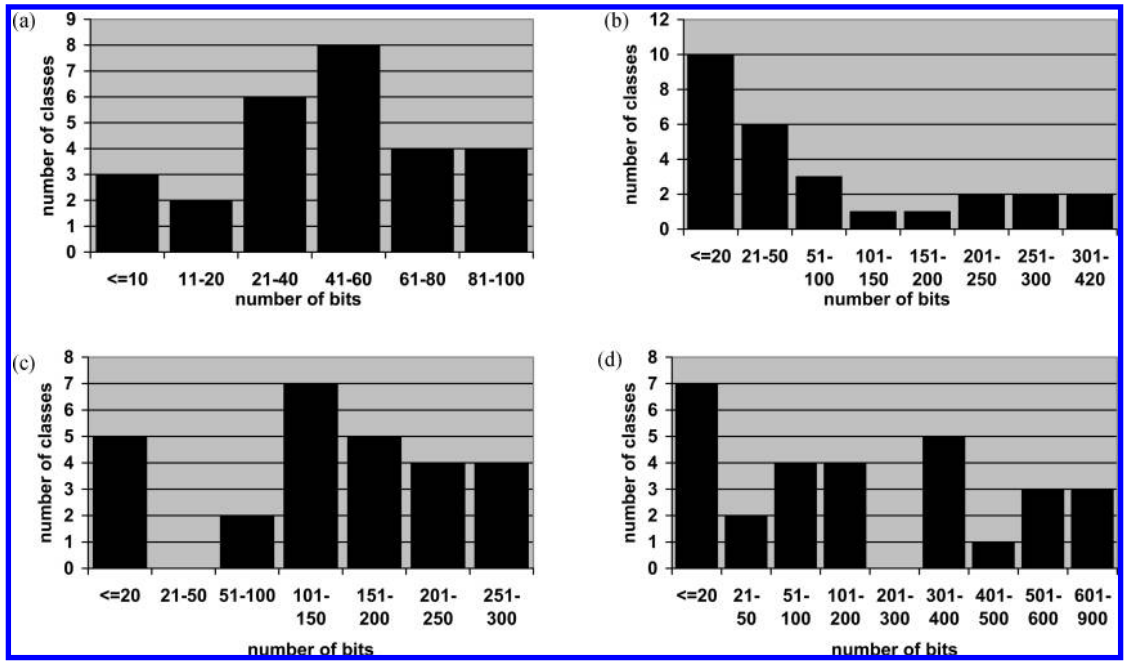

Figure 2. Bits required to achieve maximum recovery rates. Reported is the number of activity classes for which reduced fingerprints of different size performed best. Each bit interval denotes a fingerprint size range. Results are shown for database selection sets of 1000 compounds: (a) MACCS, (b) TGD, (c) ECFP, and (d) GPIDAPH3.

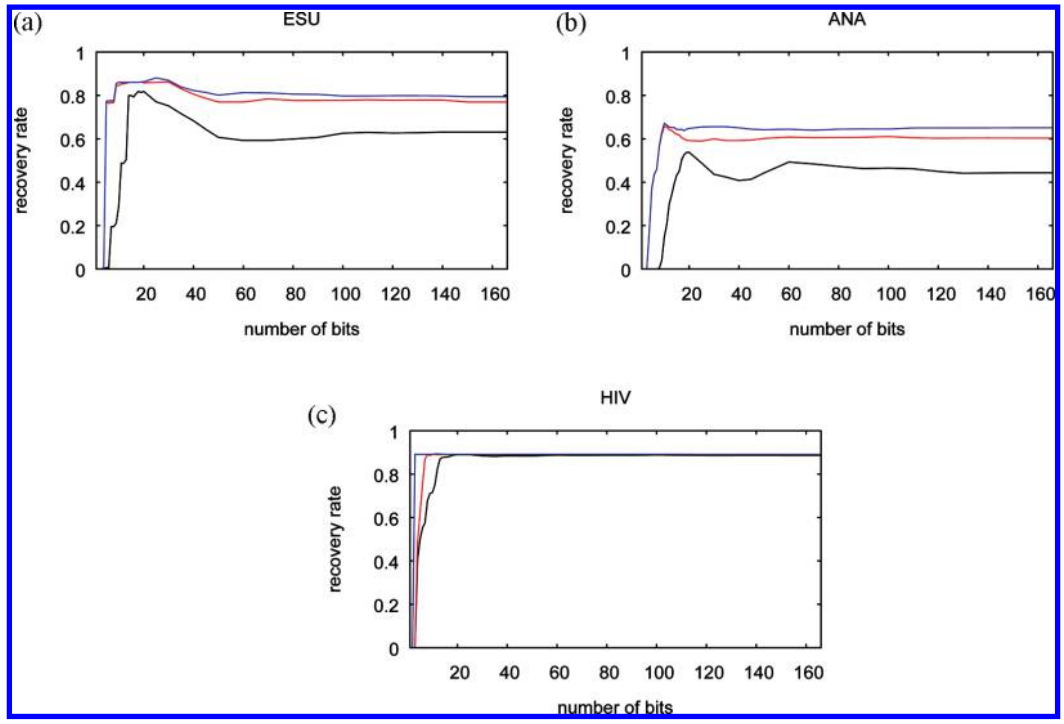

Figure 3. Compound recall curves for varying numbers of MACCS bits. For three activity classes, average recovery rates for Bayesian screening are reported for varying numbers of MACCS bit positions selected on the basis of KL divergence. Curves are shown for databases selection sets of 100 (black), 500 (red), and 1000 (blue) compounds.

divergence are combined into subsets of increasing size (starting from single bits). These fingerprint subsets are then utilized to calculate Bayesian scores for database compounds according to eq 5 as a likelihood of activity. This process is termed Bayesian screening.[12] Hence, KL-divergence analysis of bit relevance and Bayesian screening share a common theoretical foundation and can be carried out within the same methodological context. Bayesian screening has been shown to outperform conventional fingerprint search strategies in many cases.[14,15]

CALCULATIONS

Benchmark calculations were performed on 27 previously reported[12] compound activity classes containing between 30 and 159 compounds, as summarized in Table 1. From each compound class, 100 sets of 20 reference molecules each were randomly selected, and the remaining compounds were added to a subset of the ZINC7[20] of approximately 3.7 million compounds (obtained after applying in-house drug-likeness filters) that served as a background database. When calculating KL divergences all compounds in the ZINC

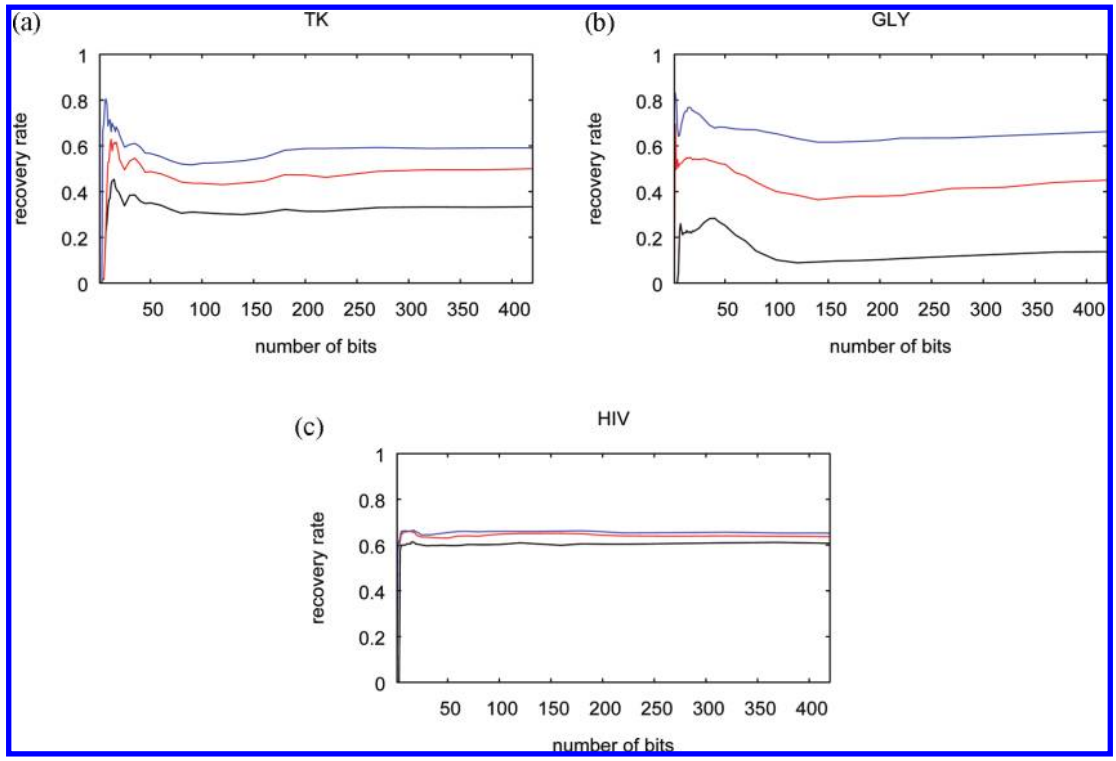

Figure 4. Compound recall curves for varying numbers of TGD bits. For three activity classes, average recovery rates for Bayesian screening are reported for varying numbers of TGD bit positions selected on the basis of KL divergence. Curves are shown for databases selection sets of 100 (black), 500 (red), and 1000 (blue) compounds.

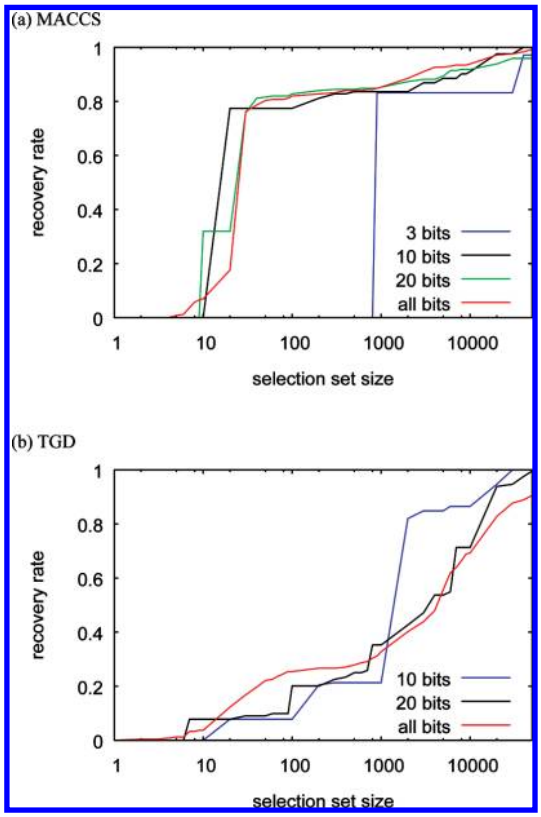

Figure 5. Screening performance of bit subsets for class HIV. For HIV inhibitors taken from PubChem, recovery rates are reported for minimal bit subsets according to Figures 3 and 4 and selection sets of varying size and compared to recovery rates of full-length fingerprints ("all bits"): (a) MACCS and (b) TGD.

database were assumed to be inactive. For each series of similarity search calculations, 100 individual trials were performed per class to obtain statistically sound results and

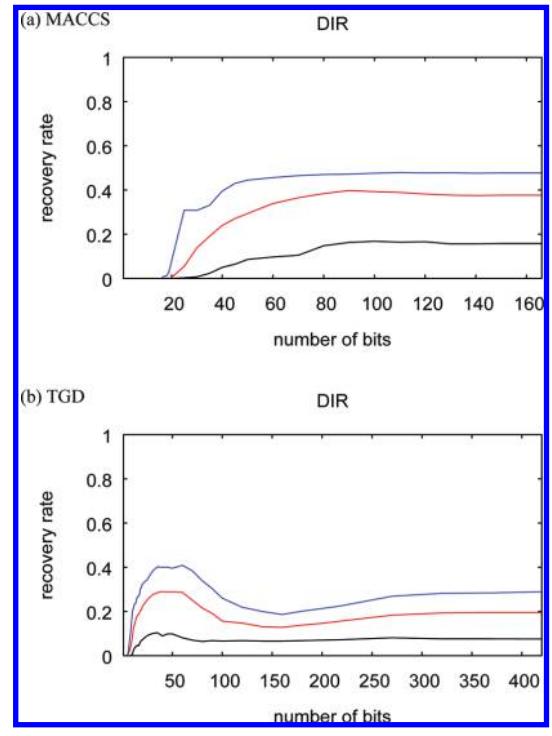

Figure 6. Compound recall curves for activity class DIR. For varying numbers of MACCS and TGD bit positions selected on the basis of KL divergence, average recovery rates for Bayesian screening are reported. Curves are shown for databases selection sets of 100 (black), 500 (red), and 1000 (blue) compounds.

averages were calculated. In addition to active compounds reported in Table 1, a total of 244 HIV protease and 27 dihydrofolate reductase inhibitors were taken from PubChem[21] to illustrate the predictive ability of minimal bit subsets on a compound set not included in the determination of bit subsets.

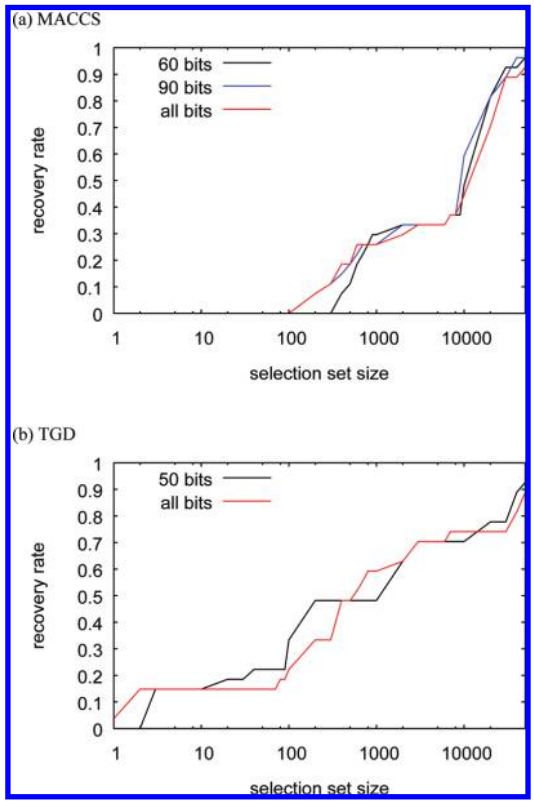

Figure 7. Screening performance of bit subsets for class DIR. For DIR inhibitors taken from PubChem, recovery rates are reported for bit subsets selected according to Figure 6 and selection sets of varying size and compared to recovery rates of full-length fingerprints ("all bits"): (a) MACCS and (b) TGD.

For our analysis, we have used four fingerprints representing different designs: MACCS[22] consisting of 166 structural keys that represent substructures or patterns consisting of 1−10 non-hydrogen atoms, the "Typed Graph Distance" (TGD) fingerprint,[23] a 2D pharmacophoric fingerprint available in the Molecular Operating Environment (MOE),[24] ECFP_4 (abbreviated ECFP),[25] an extended connectivity fingerprint with a maximum path length of four bonds around atoms implemented in Pipeline Pilot,[26] and the "Graph-Π-Donor-Acceptor-Polar-Hydrophobe-Triangle" (GPIDAPH3) fingerprint also available in MOE. TGD is a two-point topological fingerprint containing 420 bits that represent atom pairs where each atom is assigned one of seven atom types and interatomic distances are divided into 15 different bond distances. GPIDAPH3 is a three-point pharmacophore-based fingerprint consisting of 30240 bits. ECFP represents a much larger set of possible features than other fingerprints and does not have a fixed format. Rather, this type of fingerprint generates atom environmental features in a molecule-specific manner that are represented as 4-byte integers using a hashing function. The theoretically possible size of the fingerprint is four billion different features. In order to reduce the complexity of the ECFP and GPIDAPH3 fingerprints for KL divergence analysis, bits exclusively created for background database compounds were only considered if these bits were set on in more than 5% of all database compounds. Removing bits from ECFP fingerprints that exclusively occur in database but not active reference compounds has previously been shown to further increase ECFP search performance in many instances.[10]

Table 1. Compound Activity Classes

class designation	biological activity	number of compounds
	Activity Classes Assembled from the MDDR[28]	
AA2	adrenergic α-2 antagonists	35
ANA	angiotensin II-AT antagonists	45
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	glycoproteine IIb-IIa receptor antagonists	34
INO	inosine monophosphate dehydrogenase inhibitors	35
LDL	upregulator of LDL receptors	30
LIP	lipoxygenase inhibitors	41
SQS	inhibitors of squalene synthetase	42
THI	thiol protease inhibitors	34
THR	thromboxane antagonists	33
XAN	xanthine oxidase inhibitors	35
	Activity Classes Assembled from the Literature	
5HT	5-HT serotonin receptor ligands[3]	71
BEN	benzodiazepine receptor ligands[3]	59
CA	carbonic anhydrase II inhibitors[3]	159
COX	cyclooxygenase-2 inhibitors[3]	31
GRH	growth hormone secretagogue agonists[29]	100
H3	H3 antagonists[3]	52
HIV	HIV protease inhibitors[3]	48
JNK	C-jun N-terminal kinase inhibitors[29]	36
MCH	melanin-concentrating hormone[29]	30
TK	tyrosine kinase inhibitors[3]	35
	Activity Classes Assembled from Other Sources	
H1D	histamine H1 receptor antagonists[30]	36
VEG	VEGFR-2 tyrosine kinase inhibitors[31]	36

For each activity class and fingerprint, individual bit positions were ranked on the basis of KL divergence analysis relative to the background database in the order of decreasing divergence values. Then Bayesian similarity search calculations were carried out for all 27 activity classes with incrementally extended bit sets: the 20 top-ranked bits were added one-by-one (i.e., subsequent similarity search calculations were carried out using a single fingerprint bit, two bits, three bits etc.), bit positions ranked from 21 to 50 were added in subsets of five bits, bit positions ranked from 50 to 200 in subsets of 10, and bits ranked from 200 to 500 in subsets of 50. For ECFP and GPIDAPH3, additional low-scoring features beyond 500 were added in increments of 100. Reference calculations were carried out with unmodified (i.e., full-length) fingerprints. These series of calculations made it possible to evaluate similarity search performance of individual bits and bit subsets selected on the basis of KL divergence analysis. For the analysis of compound recall, recovery rates of active compounds were calculated for differently sized bit subsets and database selection sets of increasing size (i.e., 100, 500, or 1000 compounds).

RESULTS AND DISCUSSION

For all fingerprints, compound classes, and each trial, a KL divergence-based bit ranking was generated, and bit positions were prioritized on this basis and incrementally selected to produce subset fingerprints of smaller size than the original ones. These alternative fingerprint representations were then applied in Bayesian screening.

Screening Performance. The comparison of recovery rates for the best performing fingerprint bit subsets and the

Table 2. Top-Ranked MACCS Bit Positions[a]

bit	p_active	p_inactive	KL divergence	description
		(a) ESU		
40	0.861	0.003	4.64	S single bonded to O
48	0.945	0.006	4.60	heteroatom bonded to ≥3 O
39	0.861	0.004	4.36	S atom bonded to 3 O
67	0.976	0.126	1.91	S attached to heteroatom
69	0.948	0.138	1.68	QH bonded to another Q
		(b) ANA		
38	0.981	0.155	1.74	C bonded to ≥2 N and 1 C
69	0.916	0.138	1.54	QH bonded to another Q
70	0.785	0.110	1.23	N bonded to 2 heteroatoms
130	0.940	0.249	1.10	more than one bond between heteroatoms
114	0.874	0.220	0.98	CH3 attached to CH2
		(C) HIV		
139	0.878	0.024	2.89	OH group
19	0.981	0.075	2.45	7-membered ring
54	0.940	0.076	2.20	QH 3 bonds from another QH
37	0.984	0.240	1.33	C bonded to ≥1 O and ≥2 N
90	0.716	0.073	1.30	heteroatom 3 bonds from a CH2

[a] For three activity classes, the top-ranked MACCS keys, the corresponding probabilities for active and database compounds, and the resulting KL divergences are reported and the keys are described. Here "Q" means any heteroatom and "X" means any heavy atom.

Table 3. Top-Ranked TGD Bit Positions[a]

bit	p_active	p_inactive	KL divergence	description
		(a) TK		
364	0.778	0.024	2.39	D<5>D
481	0.978	0.208	1.43	A<2>A
468	0.978	0.215	1.40	A<4>D
482	0.979	0.258	1.23	A<3>A
469	0.981	0.299	1.10	A<5>D
		(b) GLY		
119	0.800	0.0003	6.04	−<15>+
644	0.743	0.001	4.32	X<15>+
539	0.714	0.006	3.05	H<15>+
434	0.571	0.002	3.02	A<15>+
551	0.829	0.023	2.67	H<12>−
		(c) HIV		
242	0.898	0.003	4.78	P<3>P
248	0.633	0.001	3.75	P<9>P
249	0.612	0.001	3.64	P<10>P
555	0.980	0.038	3.10	H<1>P
454	0.939	0.031	3.05	A<5>P

[a] For three activity classes, the top-ranked TGD keys, the corresponding probabilities for active and database compounds, and the resulting KL divergences are reported, and the pharmacophore patterns are described using different atom types (D: hydrogen bond donor, A: hydrogen bond acceptor, P: both hydrogen bond donor and acceptor, H: hydrophobic atom, "−": acid, "+": base, X: none of the above) and distance between these atoms. For example, "D<5>D" means that two donor atoms are separated by five bonds.

full-length fingerprints for selection sets of 100 and 1000 database compounds is shown in Figure 1 and reported in detail in Table S1 of the Supporting Information. For all four fingerprints, recovery rates of full-length fingerprints were always met by reduced fingerprints, and, in some cases, reduced versions performed better than the unmodified fingerprints, especially for ECFP and GPIDAPH3.

As will be discussed below in more detail, for ECFP, performance enhancements of reduced versions over the unmodified fingerprint and over MACCS and TGD were in part dramatic. Figure 2 reports the number of activity classes for which reduced fingerprints produced best recovery rates (for selection sets of 1000 database compounds) and the size distribution of best-performing fingerprints. For this purpose, reduced fingerprints were combined into bit number intervals. Figure 2a shows that for MACCS, between 11 and 100 of 166 bit positions were sufficient to produce top recovery rates for 24 classes. For the remaining three classes, fewer than 10 bits yielded highest recovery rates. These extreme cases will be further analyzed below. For TGD, shown in Figure 2b, we also observe that fewer than 20 pharmacophore features were sufficient to yield top recovery rates for 10 activity classes and that reduced TGD versions with up to 150 (of 420) bits performed best for 20 classes. For ECFP with its flexible format, our compound classes produced on average 615 and maximally 1358 features. However, as shown in Figure 2c, only between 51 and 300 features were sufficient for producing highest recovery rates in 22 cases. Also for ECFP, fewer than 20 features yielded top recovery rates for five classes. For GPIDAPH3, shown in Figure 2d, maximally 900 bits were sufficient to produce top recovery rates for all classes, which represent only ∼3% of all GPIDAPH3 bits. Furthermore, for seven activity class, fewer than 20 bits yielded top recovery rates.

As a further control for the positive effects of KL-based bit selection, random bit subsets of the same size as the best performing bit subset were selected, and average recovery rates for selection set sizes of 100 and 1000 database compounds were compared. The comparison revealed that the random bit subsets never performed better than KL-selected bits. Especially for small bits subsets, random bit subsets were inferior to the KL-selected bits. A detailed comparison of the random fingerprint subsets and the subsets selected based on KL divergence is given in Figure S1 of the Supporting Information.

Taken together, the small number of bit positions that were sufficient to achieve highest recovery rates in many cases was striking, especially taking into account that equivalent observations were made for fingerprints of distinct designs. It should also be noted that meaningful similarity searching with fingerprints of such small size using conventional search

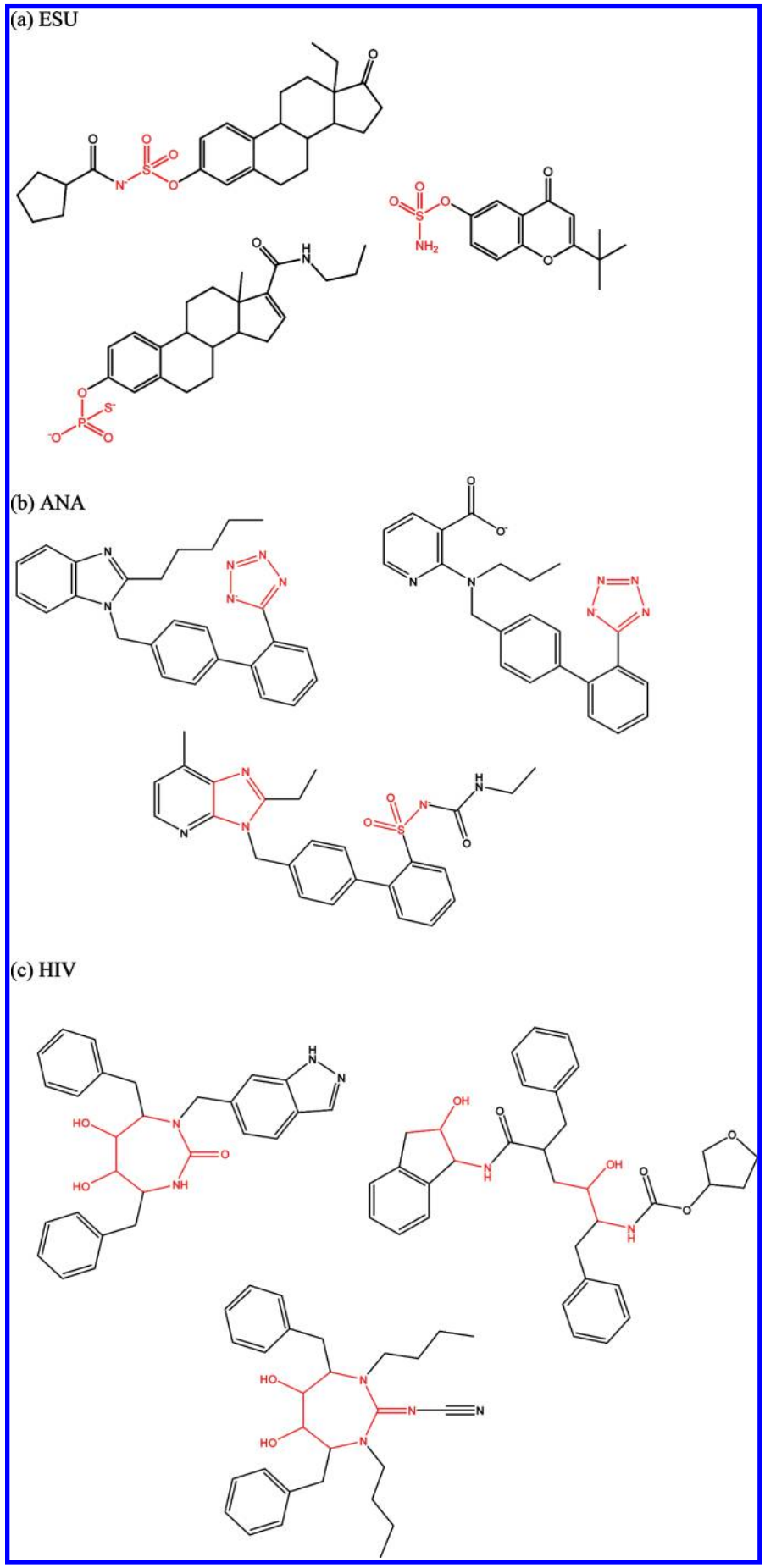

Figure 8. Mapping of top-scoring MACCS keys. For three activity classes, the top five MACCS keys (see Table 2) were mapped on representative compounds, and delineated substructures are highlighted in red.

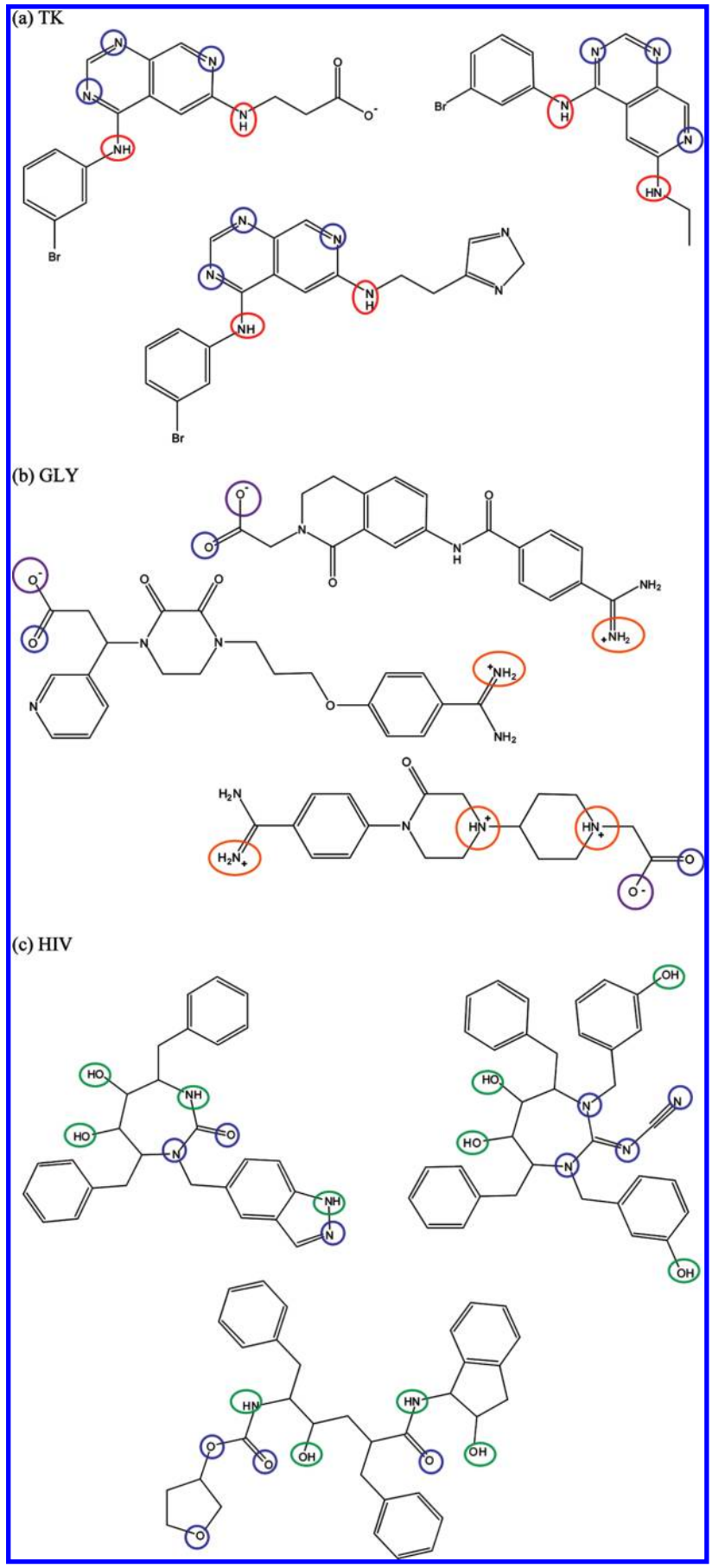

Figure 9. Mapping of top-scoring TGD keys. For three activity classes, the top five pharmacophore patterns (see Table 3) were mapped on representative compounds. Mapped pharmacophore patterns contain hydrogen bond donors (red), hydrogen bond acceptors (blue), donors/acceptors (green), negatively charged atoms (purple), and/or positively charged atoms (orange).

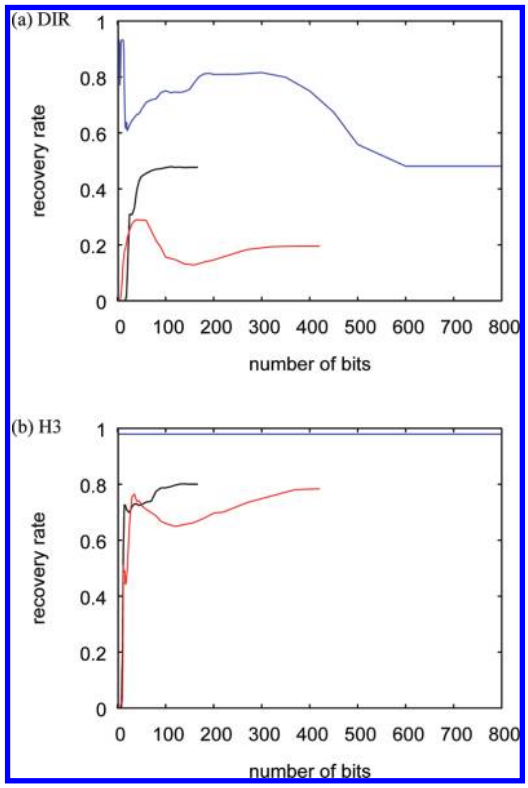

Figure 10. Fingerprint comparison. For two activity classes, recovery rates are reported for different versions of ECFP (blue), MACCS (black), and TGD (red) of increasing size. Bits were added in decreasing order of KL divergence.

methods such as the calculation of Tanimoto similarity[1] might be difficult because of statistical limitations.[27] However, as shown here, small numbers of bit positions could be effectively exploited in the context of Bayesian screening.

Structural and Pharmacophore Keys. In Figures 3 and 4, recall curves are reported for MACCS and TGD, respectively, over varying numbers of bits and for database selection sets of different size. Here we focus on compound classes for which very few bit positions were highly discriminatory and produced top recovery rates. The GPIDAPH3 fingerprint was omitted from this and the following comparisons to limit the number of representations because the results were consistent with those obtained for the other fingerprints. In all cases, the shape of the recall curves is similar for selection sets of increasing size, i.e. recovery rates of minimal fingerprints are constant or reduced when more bit positions are added. If recovery rates are reduced, the reduction occurs after addition of limited numbers of bits, and beginning with about 80 bit positions for MACCS and 150 for TGD, recovery rates remain essentially constant. In the case of HIV, top recovery rates were achieved for selection sets of 100 compounds, but in the other four cases, recovery rates increased with selection set size. However, with the exception of GLY for TGD, high recovery rates of 40%−90% were already observed for the smallest database selection sets. Thus, Figures 3 and 4 clearly illustrate the highly discriminatory nature of small subsets of MACCS and TGD bit positions for selected compound classes.

Depending on the size of the database selection sets, minimal bit subsets producing highest recovery rates for HIV in Figures 3c and 4c consisted of only three, 10, or 20 MACCS bits and 10 or 20 TGD bits. These bit subsets were also applied to search for 244 HIV inhibitors collected in PubChem added to ZINC as an additional test set using the 48 HIV inhibitors in Table 1 as reference molecules. The search results are reported in Figure 5 and further illustrate the predictive ability of minimal bit subsets. As can be seen, as few as 10 MACCS and TGD bit positions were responsible for fingerprint search performance over the entire range of selection sets.

To further investigate the bit reduction approach, a second external test set consisting of 27 dihydrofolate reductase (DIR) inhibitors was also collected from PubChem and added to ZINC. For search calculations, the 30 DIR inhibitors reported in Table 1 were used as reference molecules. Figure 6 shows the results of benchmark calculations for class DIR, and Figure 7 shows the search results for the external test set with preferred bit subsets selected according to Figure 6. Depending on the size of the database selection sets, minimal bit subsets producing highest recovery rates for DIR consisted of 60 or 90 MACCS bits and 50 TGD bits (Figure 6). The search results reported in Figure 7 illustrate that bit subsets determined based on a training set are also capable of reproducing the performance of a full-length fingerprint for external test compounds.

Tables 2 and 3 list the five top-scoring MACCS and TGD keys, respectively, for the selected activity classes together with their KL divergence. We have mapped these keys on compounds in the corresponding activity classes in order to explore their structural meaning. Figures 8 and 9 show the results for MACCS and TGD, respectively, utilizing representative active compounds. In all cases, the combination of top-scoring keys captures characteristic functional groups and ring structures (Figure 8) or hydrogen bond donor/acceptor patterns (Figure 9). These in part rather unusual structural signatures are likely to discriminate active compounds from many database molecules, consistent with the high KL divergence values of the corresponding bit positions and the strong performance of minimal fingerprint representations observed in the cases.

Extended Connectivity Fingerprint Features. As shown in Figure 1 and Table S1 of the Supporting Information, ECFP feature reduction significantly improved search performance compared to MACCS and TGD. Considering the best performing versions of these fingerprints, ECFP consistently achieved highest recovery rates and the differences were in part very large; for seven activity classes, reduced ECFP versions increased the recovery rates of MACCS and TGD by more than 40%. Figure 10a shows an example illustrating these effects. For DIR, recovery rates of ∼48% were achieved with both full-length MACCS and ECFP that generated 561 features (both fingerprints outperformed TGD in this case). However, reduction of ECFP to only a single feature increased the recovery rate to ∼95%. This feature describes a characteristic substructure shown in Figure 11a. Furthermore, Figure 10b shows another extreme example that also highlights the dramatic effect of bit reductions that we generally observed. In this case, for class H3, only four ECFP features determined the search performance of its complete ensemble of 474 atom environment strings (with a recovery rate close to 100%). For these activity classes, we also determined the top-scoring ECFP features, reported in Table 4, and mapped them onto active compounds, as shown in Figure 11. Consistent with our earlier observations for top-

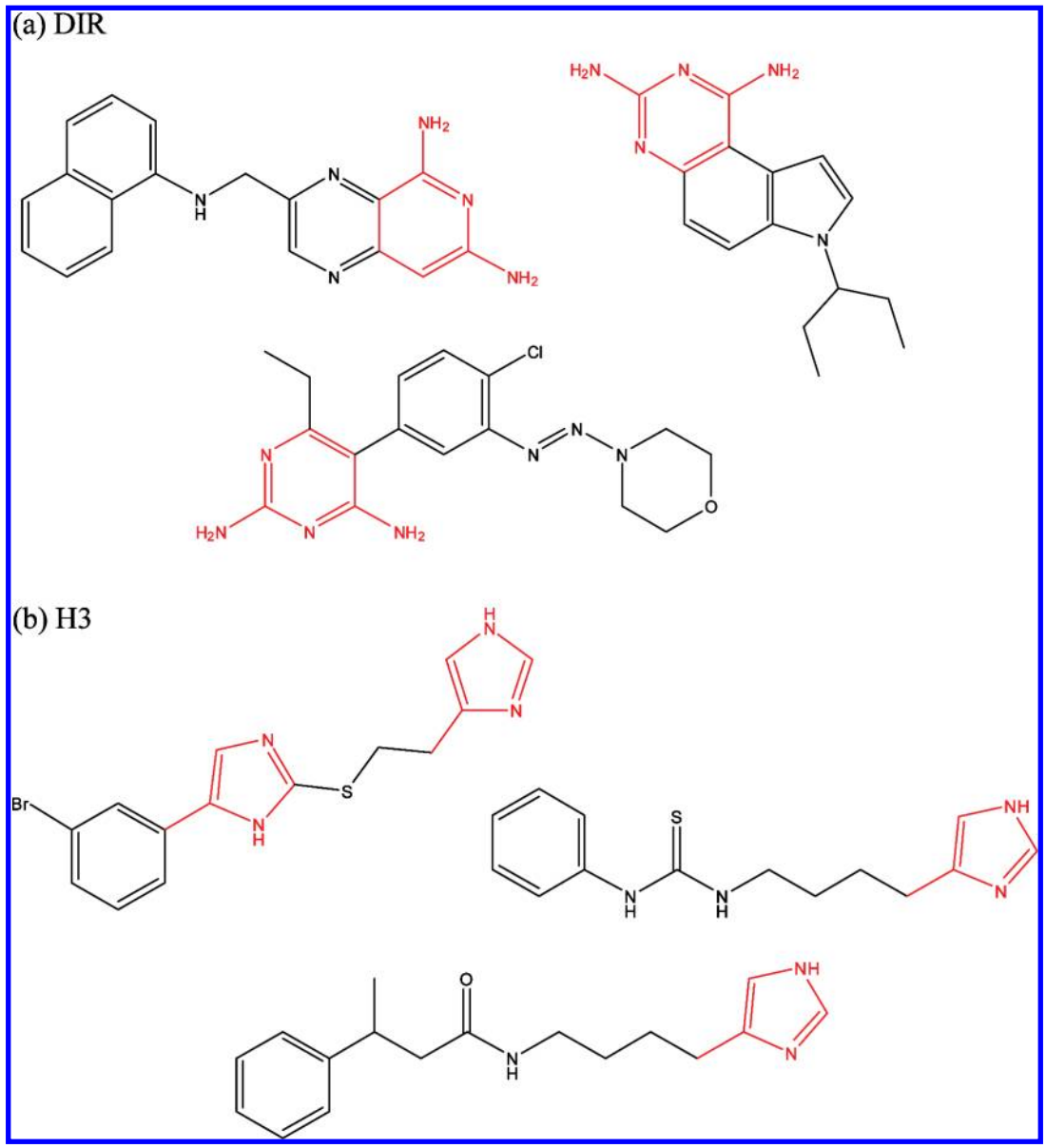

Figure 11. Mapping of top-scoring ECFP features. For two activity classes, the top five ECFP features (see Table 4) were mapped on representative compounds, and delineated substructures are highlighted in red.

Table 4. Top-Ranked ECFP Features[a]

bit	p_active	p_inactive	KL divergence	SMARTS
		(a) DIR		
−1408385607	0.903	2.0e-04	7.34	[*][c]1:[*]:n:[c](N):n:[c]:1N
−1734834311	0.903	2.0e-03	5.11	[*][c](:[*]):[c](N):n:[*]
−1735438812	0.902	3.0e-03	5.01	[*]:n:[c](N):n:[*]
−938530932	0.969	3.0e-02	3.24	[*]:[c](:[*])N
−1994733022	0.354	4.6e-05	2.89	[*]C[c]1:c:[*]:[c]([*]):n:[c]:1N
		(b) H3		
725072930	0.962	7.1e-05	9.03	[*]CC[c]1:c:n:c:n:1
1905301167	0.962	1.3e-04	8.39	[*]C[c]1:c:n:c:n:1
−750301151	0.962	1.3e-04	8.39	[*]C[c]1:c:n:c:n:1
486597464	0.981	2.8e-04	7.93	[*][c]1:c:n:c:n:1
408863435	0.981	2.8e-04	7.93	[*][c]1:c:n:c:n:1

[a] For two activity classes, the top-ranked ECFP features, the corresponding probabilities for active and database compounds, and the resulting KL divergences are reported. Features are provided as SMARTS strings.[32]

scoring MACCS and TGD keys, the five ECFP features with largest KL divergence also captured characteristic substructures in DIR and H3 compounds.

Reduced versions of ECFP also showed the overall most significant increases in search performance compared to the full-length fingerprint, as illustrated in Figure 12 for database

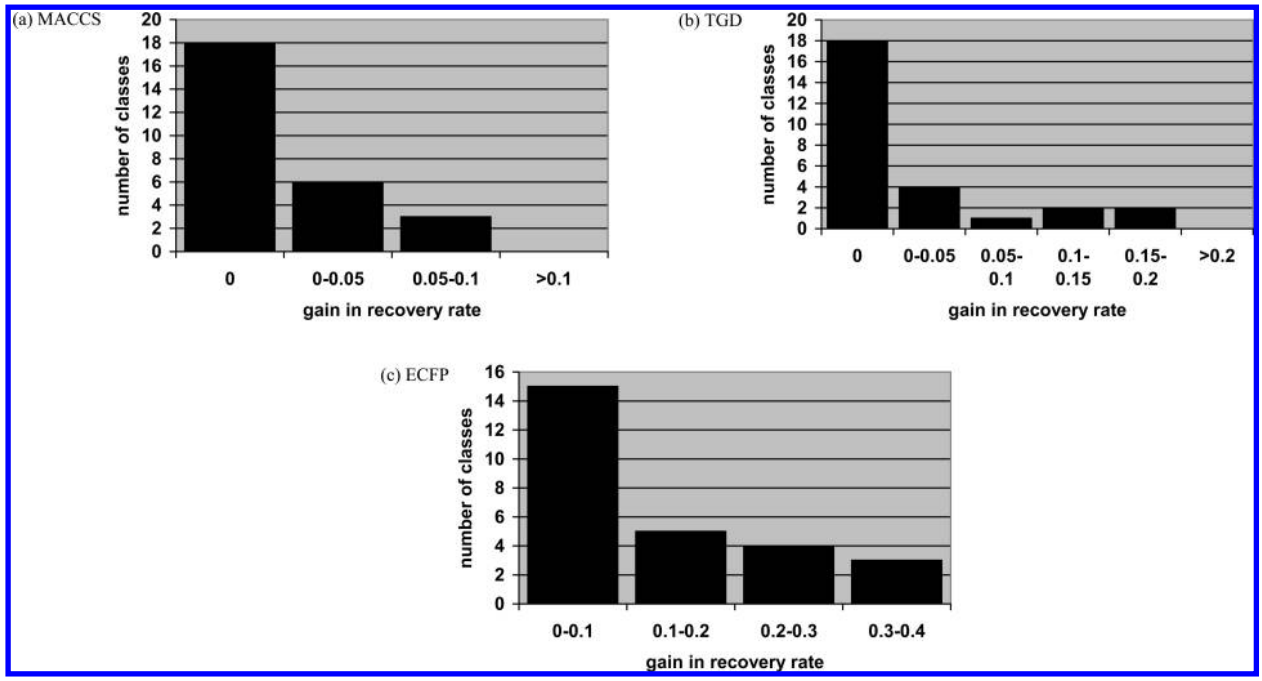

Figure 12. Increase in search performance through fingerprint reduction. For the best-performing reduced fingerprint versions of (a) MACCS, (b) TGD, and (c) ECFP, the numbers of compound classes are reported for which a certain increase in recovery rate (i.e., from 0 to 0.3−0.4) was achieved relative to full-length ECFP for database selection sets of 1000 compounds.

selection sets of 1000 compounds. For 12 of 27 classes, feature reduction increases recovery rates by 10% to 40%. Thus, reduced ECFP fingerprints are highly attractive for practical Bayesian screening applications. For MACCS and TGD, reduced bit sets match the performance of the full-length fingerprint for 18 of the 27 activity classes. Moreover, bit subsets of the remaining activity classes increase the search performance of the full-length fingerprints (up to 10% for MACCS and 20% for TGD).

For all fingerprints, we have found compound classes where only very few or single bits achieved highest recovery rates. An instructive example is the application of ECFP Bayesian screening to activity class TK. This activity class shows the highest internal similarity for ECFP fingerprints and requires only a single bit to recover all potential hits. TK compounds are found to share a large maximum common substructure, and the selected ECFP bit is a subset of this substructure.

Given these findings, we have investigated whether a relationship exists between intraclass compound similarity and bit reduction. For MACCS, the classes HIV and SQS require the lowest number of bits to achieve maximum compound recovery and activity class HIV displays highest intraclass similarity (average pairwise MACCS Tanimoto coefficient (Tc) = 0.74). By contrast, class SQS shows only low internal similarity (Tc = 0.52). Similar results were obtained for ECFP. Besides activity class TK, two additional activity classes (INO and DIR) required only a single bit to achieve the highest recovery rate. However, these two classes do not contain a representative maximum common substructure, and the intraclass similarity is significantly lower than for TK. Therefore, there was no systematic relationship between the intraclass structural similarity and the size of the selected bit subset.

CONCLUSIONS

In this study, we have introduced a Kullback−Leibler divergence method to determine the compound class-dependent significance of individual fingerprint bit positions in the context of Bayesian screening. Divergence-based ranking of bits makes it possible to reduce fingerprint representations in size by focusing on those features and the corresponding bit positions that effectively discriminate, in information-theoretic terms, between active molecules and random database compounds. For fingerprints of different design, bit reduction has been shown to produce consistent effects. In all cases, subsets of bits, often consisting of only a few positions, were shown to determine fingerprint search performance. Moreover, in many instances, reduced fingerprint representations performed better than the original unmodified fingerprints.

ACKNOWLEDGMENT

B.N. is supported by Bayer HealthCare AG, Wuppertal, Germany.

Supporting Information Available: Recovery rates for best performing bit subsets and full-length fingerprints (Supplementary Table S1) and comparison of KL-selected bit subsets and random bit subsets (Supplementary Figure S1). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046–1053.
(2) Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discovery Today* **2007**, *12*, 225–233.
(3) Xue, L.; Bajorath, J. Accurate partitioning of compounds belonging to diverse activity classes. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 757–764.
(4) Bender, A.; Mussa, Y.; Glen, R. C.; Reiling, S. Similarity searching of chemical databases using atom environment descriptors: evaluation of performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708–1718.

(5) Batista, J.; Bajorath, J. Similarity searching using compound class-specific combinations of substructures found in randomly generated molecular fragment populations. *ChemMedChem* **2008**, *3*, 67–73.
(6) Schemetulskis, N. E.; Weininger, D.; Blankley, C. J.; Yang, J. J.; Humblet, C. Stigmata: an algorithm to determine structural commonalities in diverse data sets. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 862.
(7) Xue, L.; Stahura, F. L.; Bajorath, J. Similarity search profiling reveals effects of fingerprint scaling in virtual screening. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2032–2039.
(8) Williams, C. Reverse fingerprinting, similarity searching by group fusion and fingerprint bit importance. *Mol. Diversity* **2006**, *10*, 311–333.
(9) Wang, Y.; Bajorath, J. Bit silencing in fingerprints enables the derivation of compound class-directed similarity metrics. *J. Chem. Inf. Model.* **2008**, *48*, 1754–1759.
(10) Hu, Y.; Lounkine, E.; Bajorath, J. Improving the performance of extended connectivity fingerprints through activity-oriented feature filtering and application of a bit density-dependent similarity function. *ChemMedChem* **2009**, *4*, 540–548.
(11) 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.
(12) 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.
(13) Vogt, M.; Bajorath, J. Bayesian similarity searching in high-dimensional descriptor spaces combined with Kullback-Leibler descriptor divergence analysis. *J. Chem. Inf. Model.* **2008**, *48*, 247–255.
(14) 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.
(15) Vogt, M.; Nisius, B.; Bajorath, J., Predicting the similarity search performance of fingerprints and their combination with other molecular descriptors using probabilistic and information-theoretic modeling. *Stat. Anal. Data Min*, in press.
(16) Berthold, M.; Hand, D. J. *Intelligent data analysis: An introduction*; Springer: Berlin, Heidelberg, Germany, 2007; pp 245−246.
(17) Ormerod, A.; Willett, P.; Bawden, D. Comparison of fragment-weighting schemes for substructural analysis. *QSAR* **1989**, *8*, 115–129.
(18) Hert, H.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. New methods for ligand-based virtual screening: use of datafusion and machine learning to enhance the effectiveness of similarity searching. *J. Chem. Inf. Model.* **2006**, *46*, 462–470.
(19) Kullback, S. *Information Theory and Statistics*; Dover Publications: Mineola, MN, 1997; pp 1−11.
(20) 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.
(21) PubChem. http://pubchem.ncbi.nlm.nih.gov (accessed March 5, 2009).
(22) *MACCS Structural Keys*; Symyx Technologies, Inc., Sunnyvale, CA, USA. http://www.symyx.com (accessed March 5, 2009).
(23) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical similarity using atom pair descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128–136.
(24) *Molecular Operating Environment (MOE)*; Chemical Computing Group Inc., Montreal, Quebec, Canada, H3B 3X3. http://www.chemcomp.com (accessed March 5 2009).
(25) Klon, A.; Glick, M.; Davies, J. Application of machine learning to improve the results of high-throughput docking against HIV-1 protease. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2216–2224.
(26) *Pipeline Pilot*; Accelrys Inc., San Diego, CA 92121, USA. http://accelrys.com/products/scitegic/index.html (accessed March 5, 2009).
(27) Hu, Y.; Lounkine, E.; Batista, J.; Bajorath, J. RelACCS-FP: a structural minimalist approach to fingerprint design. *Chem. Biol. Drug. Des.* **2008**, *72*, 341–349.
(28) *Molecular Drug Data Report (MDDR)*; Symyx Technologies, Inc., Sunnyvale, CA, USA. http://www.symyx.com (accessed March 5, 2009).
(29) 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.
(30) 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.
(31) Chen, X.; Liu, M.; Gilson, M. K. Binding DB: a web-accessible molecular recognition database. *Comb. Chem. High Throughput Screening* **2001**, *4*, 719–725. http://www.BindingDB.org (accessed March 5, 2009).
(32) *SMARTS*; Daylight Chemical Information Systems, Inc., Aliso Viejo, CA, USA. http://www.daylight.com (accessed March 5, 2009).

CI900087Y