DOI: 10.1002/cmdc.200700090

Introduction of a Generally Applicable Method to Estimate Retrieval of Active Molecules for Similarity Searching using Fingerprints

Martin Vogt and Jürgen Bajorath*[a]

Fingerprints are bit string representations of molecular structure and properties and are among the most widely used computational tools for similarity searching and database screening. Various fingerprint designs are available and their search performance is in general strongly dependent on the compound classes under study and the chemical characteristics of screening databases. Currently, it is not possible to predict the probability of identifying novel hits through fingerprint searching. However, for practical applications, such estimations would be very useful because one might be able, for example, to prioritize fingerprints and compound selection strategies or decide whether or not a similarity search campaign with subsequent experimental evaluation of candidate compounds would be promising at all. We have developed a method that makes it possible to predict the outcome of similarity search calculations using any type of keyed fingerprint. The methodology incorporates bit frequency distributions of reference molecules and the screening database into an information-theoretic function and determines the principally possible recall of active compounds within selection sets of varying size. We calibrate the function on diverse compound classes and accurately predict compound recovery in retrospective virtual screening trials. Furthermore, we correctly predict fingerprint search performance on two experimental high-throughput screening data sets (HTS). Our findings indicate that given a set of reference molecules, a fingerprint, and a screening database, we can readily estimate how likely it will be to retrieve active compounds, without knowledge about the distribution of potential hits in the database.

Introduction

Similarity searching using molecular fingerprints is one of the classic approaches to compound database mining.[1–4] Conventional fingerprints are bit string representations of molecular structure and properties and as such encode binary feature sequences.[2–4] The popularity of fingerprint searching is in part due to its computational efficiency and the possibility to use either single or multiple reference molecules.[2,4] A search involves pair-wise comparisons of fingerprints calculated for reference molecules and database compounds and produces a similarity-based ranking of the database. A measure of molecular similarity, fingerprint overlap is quantified using similarity coefficients such as the Jaccard or Tanimoto coefficient (Tc).[4] Fingerprint search calculations using multiple reference molecules employ data fusion techniques[5,6] or other specialized search strategies[6,7] and often produce more hits than single-template calculations,[4,8] because of the increase in molecular information guiding the search. A variety of fingerprints have been introduced that encode different 2D or 3D structural, property, or pharmacophore descriptors.[9–18] Fingerprint designs can either be keyed where each bit is associated with a specific feature[8,12] or hashed where feature combinations are mapped to overlapping bit regions.[9]

The performance of similarity search tools generally depends on the compound activity classes that are studied.[8,19] Clearly, for practical applications, it would be very helpful to predict the ability of similarity search calculations to retrieve active compounds. Then different fingerprints and compound selection strategies could be evaluated to maximize chances of success. However, such predictive approaches are currently not available.

We have investigated the possibility of estimating compound recovery rates for fingerprint search calculations, irrespective of fingerprint type. At a first glance, this appears to be a rather ambitious goal. However, we can make use of a canonical feature that is shared by fingerprints despite differences in design, complexity, and size: fingerprints are bit strings and as such binary in nature. Thus, if bit settings of different sets of compounds were transformed into binary value distributions, probabilistic modeling might be possible.

On the basis of these considerations, we decided to analyze fingerprint bit distributions using an information-theoretic function that we originally developed for compound screening using Bayesian models. Methodological details and the derivation of a novel function to predict recovery rates for fingerprint searching are described in the following section.

After introducing the methodology, we calibrate it for practical applications. For a total of 40 activity classes, we have cor-

[a] *M. Vogt, Prof. Dr. J. Bajorath*
Department of Life Science Informatics
Bonn-Aachen International Center for Information Technology
Rheinische Friedrich-Wilhelms-Universität Bonn
Dahlmannstr. 2, 53113 Bonn (Germany)
Fax: (+49) 228-2699-341
E-mail: bajorath@bit.uni-bonn.de

related predicted recovery rates with those obtained for two different fingerprint designs and three alternative search strategies. This correlation analysis has enabled us to derive linear models for prediction of recovery rates for activity classes not included into the training set. Recovery rates were well predicted for seven other compound classes and different fingerprints and search strategies. In addition, we use the newly derived function to mine two publicly available HTS data sets containing inhibitors of cathepsin B and dihydrofolate reductase to further evaluate its application potential.

Taken together, our findings suggest that using the new methodology it is readily possible to predict the outcome of fingerprint search calculations and prioritize fingerprints for given sets of reference compounds and screening databases.

Theory and methodology

First we describe the methodological basis for deriving our fingerprint function. We feel it is important to thoroughly present the theory underlying our approach. In addition, at the end of the theory section, we also give some practical guidance how to apply the function we derive for actual predictions of compound recall.

Previously, a distance function to navigate high-dimensional descriptor spaces[20] was transformed into a log-odds function[21] using principles of Bayesian modeling.[22] The distance from a center of a region in chemical space populated by active reference compounds was expressed as a log-odds estimate of activity where increasing distance from the center of the active subspace corresponded to decreasing probability of activity and vice versa.[21] For this so-called BDACCS function, values of continuous molecular property descriptors served as features. The divergence between probability functions of active molecules and database compounds could then be used to estimate the retrieval of compounds from the source database. Following this idea, we incorporated the Kullback-Leibler (KL) function[23] into BDACCS, which provides a measure for the divergence of two probability distributions and is widely applied in information theory.[23,24] The resulting KL-BDACCS method made it possible to predict recovery rates for Bayesian screening calculations.[25] This concept was adapted for studying fingerprint bit distributions and predicting the outcome of search calculations.

There are numerous distance and similarity measures for fingerprints to assess the similarity of molecules.[1] When using multiple reference compounds, individual measures of similarity can be combined using data fusion techniques like nearest neighbor searching or, alternatively, average or centroid fingerprints can be generated to produce a ranking of database compounds.[4–7] However, these approaches do not take into account the relative occurrence of certain bit positions in fingerprints of active compounds (and their potential importance) compared to the population database. For instance, if a bit position representing a specific feature is present in 70% of a class of active compounds, it might be an indicator of activity. But if this feature is also present in 90% of the background database, the probability of activity is about 3.8 times higher for the 10% of the molecules that do not possess the structural element than for those that do. Thus, incorporating the relative importance of each bit position into a similarity measure is clearly of relevance for bit string comparisons using information provided by multiple reference molecules. The importance of a bit position to contribute to a signature of active compounds can be estimated, for example, by calculating the relative frequency of occurrence in reference and database molecules. Bit frequency calculations have been carried out for instance in the context of substructural analysis methods as described by Ormerod et al.,[26] Hert et al.,[27] and in a more formal manner by Bender et al.[28] using naïve Bayesian classifiers for an atom environment fingerprint. Using bit frequency information, we formally develop a general weighting scheme for fingerprints on the basis of Bayesian statistics. More specifically, by interpreting each bit in a fingerprint as an independent random variable that adopts a binary value distribution we develop a weighting scheme for each bit position in a Bayesian framework that correctly reflects the likelihood ratios for activity, as exemplified above. This is done in analogy to the BDACCS approach where the score reflects a log-odds likelihood for activity.[21]

As fingerprints consist of binary values each bit position i can be viewed as a Bernoulli-distributed random variable v_i with $P(v_i=1)=p_i$ and $P(v_i=0)=q_i=1-p_i$ being the probabilities that the bit is set on or not, respectively. By counting the relative frequencies for each bit position for a collection of active compounds and a background database (where most compounds do not share the same activity) we can estimate the probabilities $P(v_i=1|A)=p_i^A$ and $P(v_i=1|B)=p_i^B$ for a certain bit position for active and inactive compounds. The likelihood ratio for bit position i is:[22]

$$R(v_i) \propto \frac{P_i(v_i|A)}{P_i(v_i|B)} = \begin{cases} p_i^A/p_i^B & \text{if} \quad v_i = 1 \\ q_i^A/q_i^B & \text{if} \quad v_i = 0 \end{cases} \tag{1}$$

To evaluate $R(v_i)$ it is not sufficient to only consider the bit positions that are set on; rather, bits set to zero must also be taken into account. This is computationally demanding for very long fingerprint designs such pharmacophore-type fingerprints[16,17] with millions of bits. Considering that such fingerprints are typically sparsely populated with bits that are set on, we augment the ratio by a factor such that each zero has a constant factor of one. Multiplying the right hand side by the constant factor q_i^B/q_i^A yields:

$$R(v_i) \propto \begin{cases} \left(p_i^A q_i^B\right)/\left(p_i^B q_i^A\right) & \text{if} \quad v_i = 1 \\ 1 & \text{if} \quad v_i = 0 \end{cases} \tag{2}$$

Taking logarithms produces a weight for bit position i:

$$\log R(v_i) = v_i\left(\log\frac{p_i^A}{p_i^B} + \log\frac{q_i^B}{q_i^A}\right) + const. \tag{3}$$

Due to the typically small sample size of reference molecules determining the probabilities for p_i^A as relative frequencies, it is meaningful to "smooth" the distribution by applying a form

of Laplacian correction. The equation above produces a weighting factor for each bit position. If we now consider a complete fingerprint $v=(v_i)_{i=1\ldots n}$ we can calculate $R(v)$ as $\prod_{i=1}^{n} R(v_i)$ if we assume independence of the bit positions. It should be noted that this assumption is implicit in most Bayesian 4approaches used for fingerprint-based virtual screening.[26,28] Then the log-odds approach leads to an additive weighting scheme for fingerprints:

$$\log R(v) = \sum_{i=1}^{n} v_i \left(\log \frac{p_i^A}{p_i^B} + \log \frac{q_i^B}{q_i^A} \right) \tag{4}$$

This weighting scheme is equivalent to the R4 weight for substructural analysis.[27,28] A score for each compound is thus calculated by adding up the log-odds weights for each bit set in the fingerprint. As mentioned above, when normalizing the weights by $\log \frac{q_i^B}{q_i^A}$, bits that are not set on effectively receive a zero weight and can be ignored.

The weights of bit positions are directly derived from the estimated probability distributions of active and inactive compounds. Therefore, it is possible to quantify the discriminatory power of the Bayesian weighting function by considering the divergence of the two distributions. For this purpose, the Kullback-Leibler divergence, a well-known measure in information theory to assess the difference of probability distributions,[23,24] is applied. Under the assumption of independence of bit positions, the joint distributions are $P(v|A) = \prod_{i=1}^{n} P(v_i|A)$ and $P(v|B) = \prod_{i=1}^{n} P(v_i|B)$. By elementary manipulation, the KL divergence given as a sum over all possible fingerprints

$$D(P(v|A)||P(v|B)) = \sum_{k=(0\ldots0)}^{(1\ldots1)} P(v=k|A) \log \frac{P(v=k|A)}{p(v=k|B)} \tag{5}$$

can be expressed as

$$D(P(v|A)||P(v|B)) = \sum_{i=1}^{n} p_i^A \log \frac{p_i^A}{p_i^B} + q_i^A \log \frac{q_i^A}{q_i^B} \tag{6}$$

Thus, the KL divergence directly corresponds to the expected value for the score of an active compound by

$$\begin{aligned} E[\log R(v|A)] &= \sum_{i=1}^{n} p_i^A \left(\log \tfrac{p_i^A}{p_i^B} + \log \tfrac{q_i^B}{q_i^A} \right) \\ &= D(P(v|A)||P(v|B)) + \sum_{i=1}^{n} \log \tfrac{q_i^B}{q_i^A} \end{aligned} \tag{7}$$

We name this function FP-KL-BDACCS (with FP abbreviating FingerPrint). If the screening database contains a small—yet unknown—number of compounds having similar activity as the reference molecules the KL divergence correlates with the percentage of active compounds among database compounds producing best scores according to the weighting scheme. Therefore, given a reference set of active compounds whose chemical features reflected by fingerprint settings are similar to those of potential hits we are able to predict the recovery rate of active compounds.

This methodology is in principle applicable to any fingerprint design but from a theoretical point of view, there are two notable exceptions. First we consider value range encoding fingerprints,[13,15] where the value range of a single chemical descriptor is encoded over a segment of bits. Then either a single bit or a number of consecutive bits are set to account for the value range, which would violate the assumption of independence of bit positions. Of course, this problem might be circumvented by estimating the distribution of the bit patterns accounting for each feature instead of considering single bit positions. The second exception is the use of hashed fingerprint designs such as the Daylight fingerprints,[12] where molecular properties or connectivity patterns are mapped to overlapping subsets of bit positions through a hashing function, which also compromises the assumption of bit independence.

Compared to Bayesian modeling of continuous value distributions, the fingerprint bit settings provide a significant advantage. The binary distribution capturing bits settings is discrete and does not require making the critical assumption that feature values follow a normal distribution.

Practical prediction of compound recall

To complement the theory we provide a practical step-by-step guide to apply our method for the prediction of compound recovery rates for a given fingerprint, a given set of activity classes, and a compound screening database. This should make it easily possible to implement the method for practical applications.

1) Fingerprint calculation and bit frequency determination:

For each compound in a reference set of an activity class A and the compound database B, the fingerprint is calculated. For each bit in the fingerprint, the relative frequencies of occurrence

$$p_i^A = \frac{T_i^A + \frac{T_i^B}{N^B}}{N^A + 1} \text{ and } p_i^B = \frac{T_i^B + \frac{T_i^A}{N^A}}{N^B + 1} \tag{8}$$

are determined. Here T_i^A and T_i^B are the number of compounds in the activity class A and the database B where bit i of the fingerprint is set on; N^A and N^B are the total number of compounds, respectively. Note that by adding a Laplacian correction to the estimated probabilities the problem of either p_i^A or p_i^B being zero can be avoided in the calculation of probability ratios because bits never set on in either the activity class or the database can be ignored.

2) Using these bit frequency values the KL function is calculated:

$$D(P(v|A)||P(v|B)) = \sum_{i=1}^{n} p_i^A \log \frac{p_i^A}{p_i^B} + (1-p_i^A) \log \frac{1-p_i^A}{1-p_i^B} \tag{9}$$

When the KL function is calculated for a number of activity classes, a given fingerprint, and a screening database, high

values for activity classes will indicate better recall rates compared to classes with lower values.

3) Calibrating the divergence function:

To quantify this relationship and compare the recall potential of different fingerprints with each other, a correlation between the logarithm of the KL function and the measured recall rate of virtual screening trials is established. Practically, for different activity classes, a number of virtual screening trials are performed where a subset is selected as reference molecules and the remaining compounds are added to the database as potential hits. For the reference set, the KL function is calculated and a similarity search is performed. Recall rates are determined as the number of recovered active compounds divided by the total number of active compounds added to the database. Thus, two values, namely KL function and recovery rate, are determined for each trial.

4) Linear regression model:

As the logarithm of the KL function correlates with the recall rate, these rates are plotted against the logarithm of the KL function and a linear function is fitted to the data points that captures the relationship between recall potential and the KL function.

5) Predicting compound recall:

For a prospective similarity search application, one requires a fingerprint, a set of known active molecules, and a screening database. Then the KL divergence is calculated as described above and the graph of the linear regression model generated in the previous step is used to "look up" the predicted recovery rate. This rate provides an estimate for the chances to succeed with a similarity search if the screening database contains molecules having similar activity (that is, potential hits).

Calculations

Reference calculations were carried out using two representative 2D fingerprints implemented in the Molecular Operating Environment (MOE)[29] that differ in design and descriptor composition. One is an implementation of atom pair descriptors[30] as a three-point pharmacophore-type fingerprint based on 2D molecular graphs. This fingerprint, termed TGT, consists of up to 2600 bits and distinguishes four atom types and six different graph distance ranges. The other is a structural fragment fingerprint, termed MACCS, and consists of 166 bits that encode a set of publicly available MDL structural keys, representing substructures consisting of one to ten nonhydrogen atoms.[11]

Three different search strategies for multiple reference compounds were applied; the centroid technique,[6] 5-NN nearest neighbor calculations,[6] and the R4 bit weighting scheme from substructural analysis.[27] A centroid fingerprint of an activity class is calculated by averaging the individual fingerprints of all reference compounds. In 5-NN search calculations, similarity scores are averaged for each test compound over the five most similar reference molecules. Thus, centroid and 5-NN search strategies operate at different levels, either fingerprint generation (centroid) or similarity scoring (5-NN). Moreover, R4 weighting differs from both strategies in that it not only considers relative bit frequencies of active reference compounds but also the bit distribution of the database.

Calculations were carried out on an in-house generated subset of the ZINC[31] database that contained ~1.44 million compounds having unique 2D graphs. Forty different compound activity classes were used to study the relationship between KL divergence and compound recovery rates FP-KL-BDACCS. The composition and origins of these activity classes are summarized in Table 1.

For each activity class, 100 randomly selected sets of ten active compounds were used as reference molecules for 100 independent search trials and in each case, the remaining four to 149 active molecules were added to the background database as potential hits. All ZINC compounds were considered decoys. The FP-KL-BDACCS scoring ranges of the 100 top scoring compounds were identified for each calculation, the number of active molecules falling into these scoring ranges determined, and recovery rates calculated.

The KL divergence was then calculated for each activity class from the relative frequencies of each fingerprint bit position for reference and database molecules. Based on these data calculated for every combination of a fingerprint and search strategy, linear regression models were derived for average recovery rates relative to the logarithm of the corresponding KL divergence. These linear functions were then used to predict recovery rates from calculated KL divergence for seven other activity classes not included in the calibration set, as summarized in Table 2. For these classes, 100 independent virtual screening trials were carried out using randomly chosen sets of ten compounds and average recovery rates for database selection sets of 100 compounds were determined.

Two HTS data set were subjected to recovery rate predictions and screening trials. The first one was generated at the Penn Center for Molecular Discovery at the University of Pennsylvania[32] using assays for inhibitors of the thiol protease cathepsin B and is available in PubChem.[33] It consists of 63 332 compounds and contains 40 hits with IC_{50} values ranging from 46 nM to 46 μM. The second HTS data set consists of 50 000 molecules screened against dihydrofolate reductase (DHFR)[34] and contains 32 confirmed competitive DHFR inhibitors with K_i values ranging from 26 nM to 11 μM.[34,35] The FP-KL-BDACCS function was recalibrated for each HTS data set using the 40 training classes reported in Table 1. Fingerprint search calculations were carried out in analogy to the activity class trials, that is, 100 sets of ten hits each were randomly selected from the HTS data and used as reference molecules for mining the data sets and searching for the remaining hits. These calculations were carried out using both TGT and MACCS fingerprints and the R4 search strategy only because no substantial differences in activity class predictions using different search strategies were observed, as further discussed below. Hit recovery was predicted from KL divergence and compared to the fingerprint search results.

Table 1. Compound classes used to calibrate the FP-KL-BDACCS function.

Class Designation	Biological Activity	No. Compds
Activity classes assembled from the MDDR[37]		
AA2	adrenergic α-2 agonists	35
ACA	ACAT inhibitors	21
ANA	angiotensin II-AT antagonists	45
ARO	aromatase inhibitors	24
CAL	calpain inhibitors	28
CHO	cholesterol esterase inhibitors	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
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
Activity classes assembled from the literature[7,36,38]		
BEN	benzodiazepine receptor ligands[38]	59
BLC	β lactamase inhibitors[7]	14
CA	carbonic anhydrase II inhibitors[38]	159
CAA	calcium antagonists[7]	18
COX	cyclooxygenase-2 inhibitors[38]	31
D2 A	dopamine D2 antagonists[7]	14
GHS	growth hormone secretagogue agonists[36]	14
GRH	gonadotropin releasing hormone agonists[36]	100
H3	H3 antagonists[38]	52
HIV	HIV protease inhibitors[38]	48
JNK	C-jun N-terminal kinase inhibitors[36]	36
MCH	melanin-concentrating hormone[36]	30
PAR	PPAR γ agonists[7]	16
PKC	protein kinase C inhibitors[7]	15
RTI	reverse transcriptase inhibitors[7]	15
TK	tyrosine kinase inhibitors[38]	35
Activity classes assembled from other sources[39–42]		
ADR	β-receptor anti-adrenergics[39]	16
CDK1	cyclin-dependent kinase 1 inhibitors[41]	22
CDK2	cyclin-dependent kinase 2 inhibitors[41]	24
FAC	factor Xa inhibitors[42]	14
GLU	glucocorticoid analogues[39]	14
H1D	histamine H1 receptor antagonists[40]	36
M2	muscarinic M2 receptor antagonists[40]	20
NET	norepinephrine transporter inhibitors[40]	21
VEG	VEGFR-2 tyrosine kinase inhibitors[41]	36

Table 2. Compound activity classes for the prediction of recovery rates for fingerprint searching.[a]

Class Designation	Biological Activity	No. Compds	avTc
5HT	5-HT serotonin receptor ligands[38]	71	0.67
ADR	β-receptor anti-adrenergics[39]	16	0.74
ARI	aldose reductase inhibitors[34]	24	0.47
BEN	benzodiazepine receptor ligands[38]	59	0.69
DD1	dopamine D1 agonists[34]	30	0.57
KAP	κ agonists[34]	25	0.57
XAN	xanthine oxidase inhibitors[34]	35	0.56

[a] In order to assess intra-class structural diversity, average values of the Tanimoto coefficient[1] (avTc) were calculated for pair-wise comparison of compounds using a set of 166 publicly available MACCS structural keys.[11]

Correlation of KL divergence and observed recovery rates

For two fingerprints and three search strategies, we studied the relationship between KL divergence and compound recovery rates observed in test calculations on the 40 different activity classes summarized in Table 1. We used a large number of activity classes to calibrate the FP-KL-BDACCS function that spanned almost the entire range of possible recovery rates for the fingerprints and search strategies we investigated, as shown in Figure 1. Recovery rates significantly varied among the 40 activity classes. Given its design, the R4 search strategy is most suitable for establishing a correlation between compound recall and KL divergence. However, we also observed significant correlation for the alternative search strategies. The chosen search strategy in part influenced the degree of observed correlation between the logarithm of KL divergence and compound recovery rates. Whereas the R4 bit frequency weighting scheme and 5-NN nearest neighbor approach produced similar correlations with both fingerprints, the centroid (fingerprint averaging) method gave a better correlation with MACCS than TGT, where recovery rates where lower. The 5-NN and R4 search strategies produced generally higher recovery rates than the centroid approach. The MACCS fingerprint produced slightly better correlations than TGT and MACCS/R4 was the preferred combination of a fingerprint and search strategy. However, the observed differences in correlation between different combinations of fingerprints and search strategies were overall not significant. For all six series of calculations, approximate linear relationships between recovery rates and the logarithm of the KL divergence became apparent. Linear regression models were derived that produced correlation coefficients ranging from 0.60 (TGT/centroid) to 0.75 (MACCS/R4) and average prediction errors of ~16% to ~18%. These linear regression models were used to predict recovery rates for other compound classes.

Prediction of recovery rates

For the prediction of recovery rates for similarity searching using fingerprints we randomly selected seven compound activity classes, summarized in Table 2. Our only selection criteri-

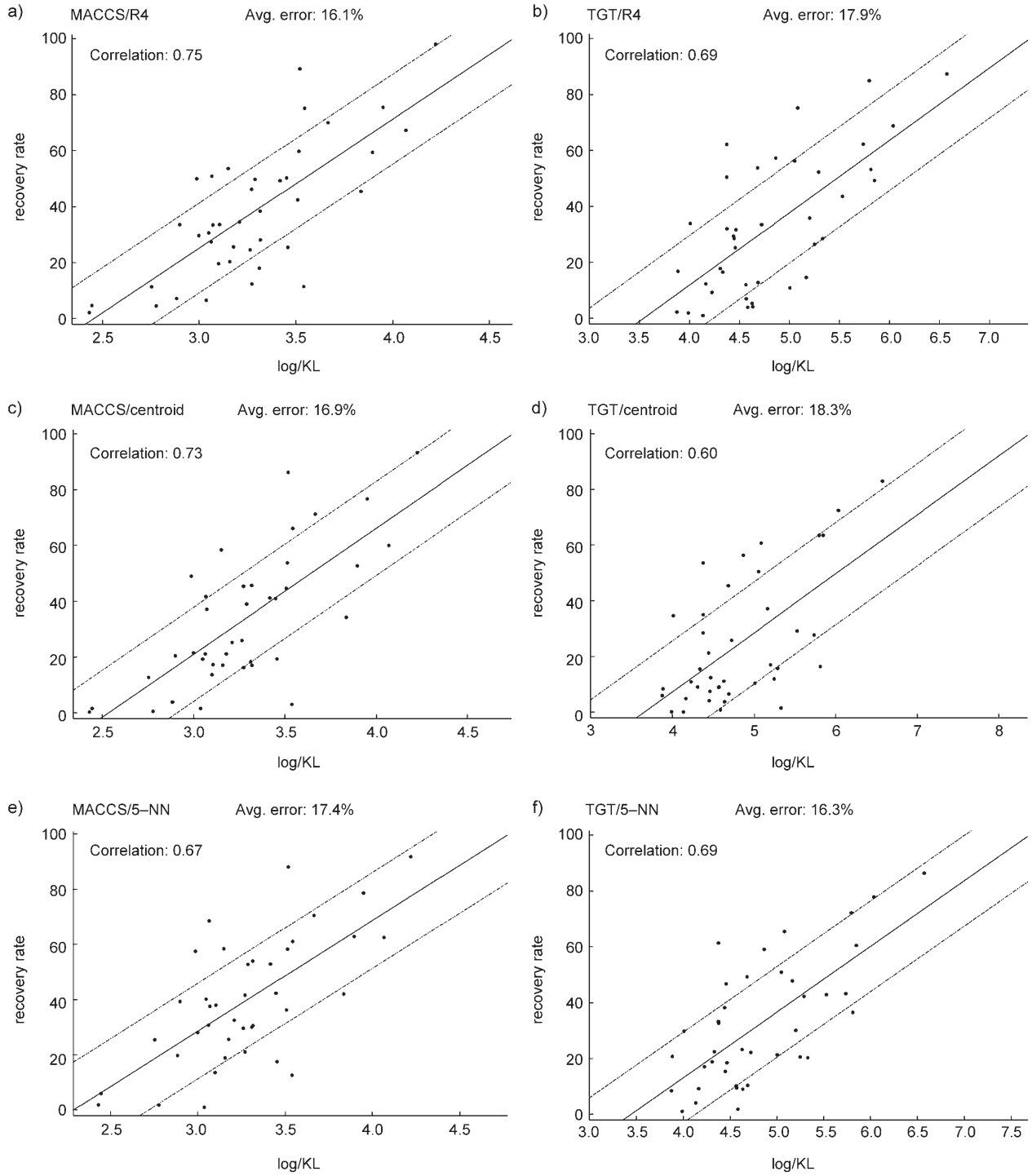

Figure 1. Linear regression analysis of training data. For 40 different compound activity classes, average recovery rates from 100 individual trials were calculated and plotted against the logarithm of the corresponding average KL divergence for the MACCS and TGT fingerprint using the R4, centroid, and 5-NN search strategies. The solid lines represent the linear regression curves and the dotted lines show the average error margins.

on was that they had to present "easy" (that is, high recovery rates) and "difficult" (low recovery rates) cases for fingerprint search calculations. This was evaluated by initial fingerprint test calculations using a few randomly selected sets of reference molecules. As indicated by averaged pair-wise MACCS Tc calculations, these compound sets had considerable intraclass structural diversity, as shown in Table 2. To produce statistically sound samples taking the influence of reference molecule set composition on recovery rates into account we carried out 100 individual trials for each combination of a selected activity class, fingerprint, and search strategy. Representative results of individual trials for a fingerprint/search strategy combination are shown in Figure 2. The Figure reveals different degrees of clustering of activity class predictions using 100 different refer-

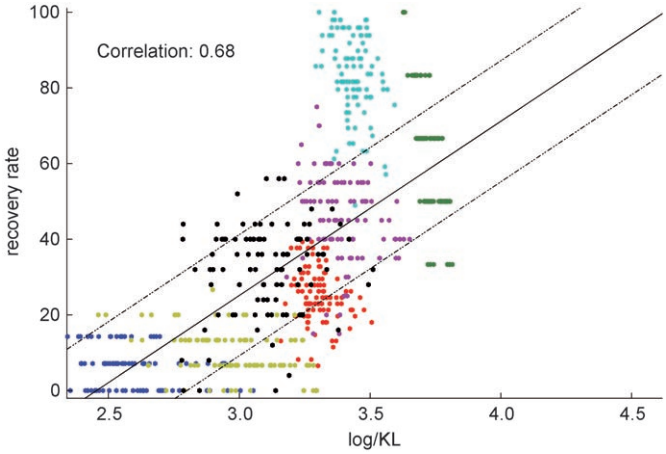

Figure 2. Prediction of recovery rates for different sets of reference molecules. As a representative example, for the seven activity classes in Table 2, the recovery rate for each of the 100 individual trials is plotted against the logarithm of the KL divergence for the TGT fingerprint and the R4 search strategy. The regression model is displayed according to Figure 1b and the activity classes are color-coded as follows: yellow, KAP; red, 5HT; cyan, BEN; green, ADR; magenta, DD1; black, XAN; blue, ARI.

ence sets covering a range from low to high recovery rates. From individual trials, average recovery rates were calculated and their correlation with KL divergence is presented in Figure 3. Compound recovery rates were predicted and measured for database selection sets of 100 compounds. A comparison of predicted and observed recovery rates is reported in Table 3. For each activity class, recovery rates varied when different combinations of fingerprints and search strategies were used, as one would expect. However, regardless of such differences, and with only very few exceptions (class BEN and MACCS), our linear regression models were able to predict observed recovery rates from KL divergence of bit settings with overall reasonable to high accuracy. For each fingerprint/search combination, meaningful predictions were obtained, irrespective of search performance. As shown in Table 3, for 15 of the 42 individual similarity search trials, average prediction errors were within 5% and for 28 trials within 10%. Moreover, with the exception of the MACCS/R4 calculation on class BEN, the magnitude of all recovery rates was well estimated. As further discussed below, this is of particular relevance for practical applications and prospective recovery rate predictions. Taken together, these findings suggested that we were able to successfully predict the outcome of fingerprint search calculations irrespective of a chosen fingerprint type or search strategy.

Mining HTS data

Having demonstrated the ability to predict compound recovery rates for activity classes added to ZINC, we next analyzed two HTS sets because these experimental data sets more closely mimic screening conditions than typical benchmark scenarios. We recalibrated the FP-KL-BDACCS function for each of these HTS sets, predicted compound retrieval, and then carried out systematic similarity search calculations in analogy to the different activity classes. For the prediction of recovery rates, we also selected the top 100 molecules from the cathepsin B and DHFR HTS sets, respectively, corresponding to 0.17% and 0.20% of the total data set size. The linear regression models and the results of the predictions and search calculations are shown in Figure 4. Similarity searching using MACCS produced average recovery rates of 16.6% (DHFR) and 18.3% (cathepsin B) and the corresponding recovery rates for TGT were 15.0% and 28.0%, respectively. Thus, recovery rates were relatively low for both fingerprints on these HTS data sets. This could be attributed to the fact that the cathepsin B and DHFR hit sets were structurally highly diverse, yielding average Tc values in pair-wise comparisons using MACCS of 0.46 and 0.38, respectively. Representative examples illustrating the structural diversity of hits are shown in Figure 5. However, despite the presence of overall low compound recall, applying the FP-KL-BDACCS function, we were able to predict these recovery rates with high accuracy, at least in three of four cases. The largest average prediction error was ~12% for TGT/DHFR where we overestimated compound recovery, predicting a rate of 26.8%. For the MACCS/cathepsin B combination, the prediction error was less than 7% and for TGT/cathepsin B and MACCS/DHFR, recovery rates were almost exactly predicted, yielding observed versus estimated rates of 28.0% versus 28.3% and 16.6% versus 14.7%, respectively. Thus, the level of accuracy of estimating recovery rates for experimental screening sets was encouraging and further supported the conclusions drawn from the analysis of diverse compound activity classes. Taken together, our results indicated that compound recovery by fingerprint similarity searching could, in general, be well predicted with the newly derived FP-KL-BDACCS function.

Discussion

Binary value distributions resulting from fingerprint bit settings provided an excellent starting point for the extension of the BDACCS approach to predict compound recovery for fingerprint calculations. For Bayesian modeling, we needed to assume that continuous descriptor value distributions follow a Gaussian distribution. By contrast, bit settings represent a Bernoulli distribution and no further assumptions are required. The FP-KL-BDACCS method can be applied to any fingerprint design where bits are not hashed or set in a cumulative manner to encode value ranges. The only assumption of the approach is that bit settings are independent of each other. For some bit settings of keyed fingerprints, this is an approximation, for example, when encoded structural descriptors or pharmacophore patterns overlap. We established that the divergence between bit distributions of reference and database compounds is proportional to the score expected for active compounds and that it correlates with recovery rates. Originally, we implemented the KL-BDACCS methodology to predict recovery rates for Bayesian modeling and screening and validated it on various compound activity classes.[25] Bayesian modeling is a rather specialized approach and can at present not be compared to the important role fingerprint search calculations have traditionally played and continue to play in vir-

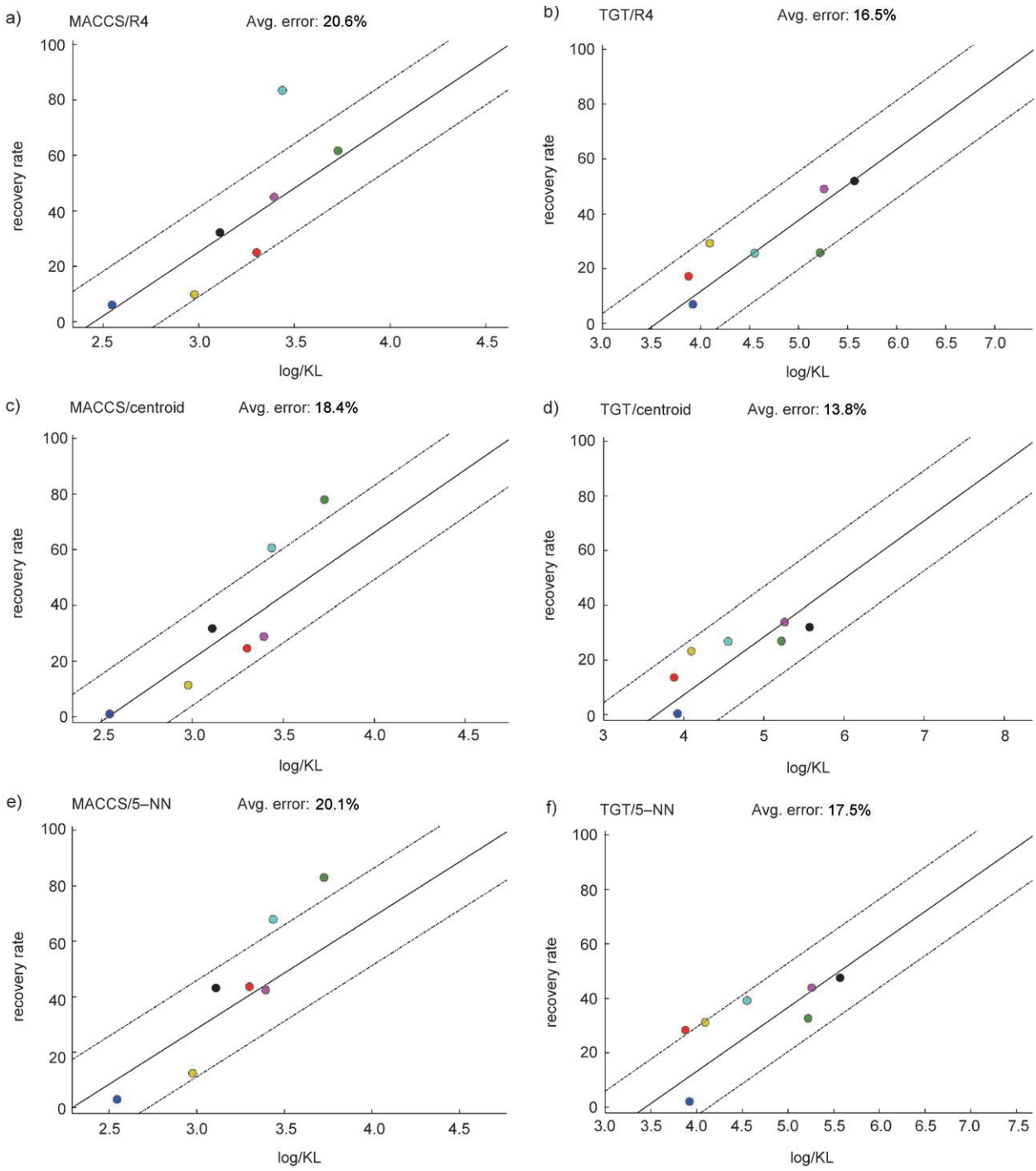

Figure 3. Prediction of average recovery rates. The presentation is according to Figure 1. For seven activity classes, recovery rates were predicted from KL divergence and the regression models. For 100 trials the average of the achieved recovery rates is shown for database selection sets of 100 compounds. The activity classes are color-coded according to Figure 2.

tual screening. Therefore, the extension of the KL-BDACCS method to generally predict recovery rates for fingerprint searching has been crucial step forward for us. In this study, we have derived the FP-KL-BDACCS function and calibrated it for different screening databases. For calibration, a large number of compound classes were used but considering the results of linear regression analysis, fewer classes would also be sufficient. We then demonstrated the ability of the method to accurately predict recovery rates for different activity classes and, in particular, HTS data sets. These findings suggest that FP-KL-BDACCS calculations can indeed be widely applied as a predictive or diagnostic tool for fingerprint searching using multiple reference compounds. The only other approach that is currently available to analyze fingerprint search calculations from a more general perspective was previously developed by us,[36] but focused on a different concept, not the estimation of

Table 3. Predicted and observed recovery rates for the seven test classes.[a]

Class Designation	MACCS Recovery Rate (Predicted rates)			TGT Recovery Rate (Predicted Rates)		
	R4	Centroid	5-NN	R4	Centroid	5-NN
5HT	25.0%	24.6%	43.6%	17.2%	13.7%	28.3%
	(39.1%)	(34.5%)	(40.6%)	(8.5%)	(4.7%)	(10.3%)
ADR	61.7%	78.0%	83.0%	25.8%	27.0%	32.7%
	(58.6%)	(53.7%)	(57.6%)	(43.3%)	(33.1%)	(41.8%)
ARI	6.1%	1.1%	3.0%	7.0%	0.4%	2.1%
	(4.3%)	(0.4%)	(10.3%)	(9.7%)	(5.7%)	(11.3%)
BEN	83.4%	60.6%	67.9%	25.7%	26.8%	39.2%
	(45.3%)	(40.6%)	(46.0%)	(26.0%)	(19.0%)	(26.1%)
DD1	44.9%	28.9%	42.5%	49.0%	33.9%	43.9%
	(43.3%)	(38.7%)	(44.3%)	(44.3%)	(34.0%)	(42.8%)
KAP	9.9%	11.4%	12.4%	29.3%	23.3%	31.2%
	(24.1%)	(19.9%)	(27.6%)	(14.1%)	(9.3%)	(15.4%)
XAN	32.2%	31.8%	43.2%	52.0%	32.1%	47.6%
	(30.2%)	(25.9%)	(32.9%)	(52.4%)	(40.6%)	(50.1%)

[a] Average recovery rates were predicted from KL divergence and determined over 100 individual trials for selection sets of 100 database compounds. Predictions are reported in parentheses.

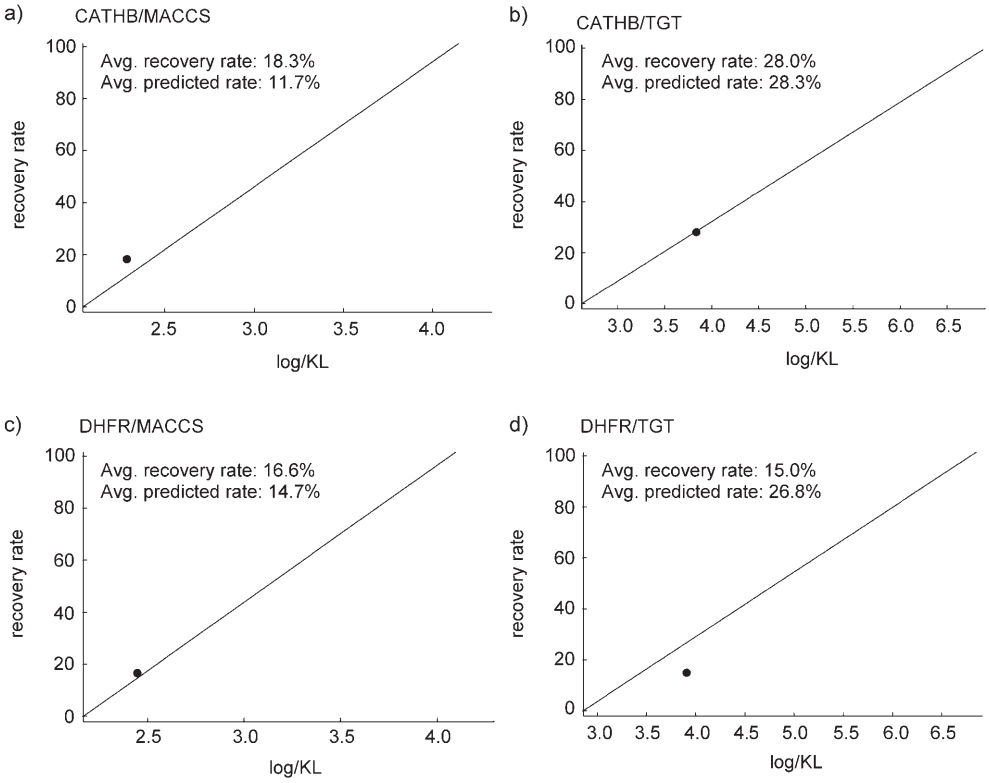

Figure 4. Recovery rates for HTS data sets. For analyzing the cathepsin B (CATHB, panels a and b) and dihydrofolate reductase (DHFR, c and d), the MACCS and TGT fingerprints were used in combination with the R4 search strategy. Average recovery rates for the top 100 ranked compounds using 100 individual calculations with different sets of 10 reference molecules were calculated and predicted. The curves represent the linear regression models obtained by calibrating the FP-KL-BDACCS function against the panel of 40 training set classes and the HTS data sets as background. Black dots indicate the averaged measured recovery rates.

recovery rates. For various compound classes and fingerprints, Tc value ranges were determined where active molecules could be successfully detected. Then these ranges were compared in histograms to the Tc value distribution obtained for database compounds. This graphical analysis revealed cases where very large numbers of database compounds were recognized within the same Tc range as potential hits. It was concluded that in such cases, fingerprint searching was a difficult, if not hopeless, exercise.[36] This technique made it possible to distinguish between favorable and unfavorable cases for fingerprint searching but could not estimate recovery rates or the retrieval of potential hits. Importantly, for the latter purpose, FP-KL-BDACCS can be practically applied in a prospective manner. For any set of active reference compounds and a given screening database, we can calculate prospective recovery rates from fingerprint bit settings, regardless of whether the database contains compounds having similar activity or not. If predicted recovery rates are low for a given fingerprint, the probability of identifying any novel hits is also low. If prospective recovery rates are high, a fingerprint search would be promising and novel hits could be identified if present in the screening database. Alternative fingerprints can be tested to select those that maximize predicted compound recall.

Conclusions

We have introduced a first methodology to predict retrieval of active compounds for similarity searching using keyed fingerprints, estimate the outcome of fingerprint searching in a prospective manner, and prioritize fingerprint types for given search situations. Theoretical foundations of the approach and the underlying theory have been discussed. Applying the FP-KL-BDACCS function, recovery rates could be accurately predicted for various compound classes and experimental screening sets. On the basis of these findings, we propose that this approach should be widely applicable as a diagnostic and predictive tool

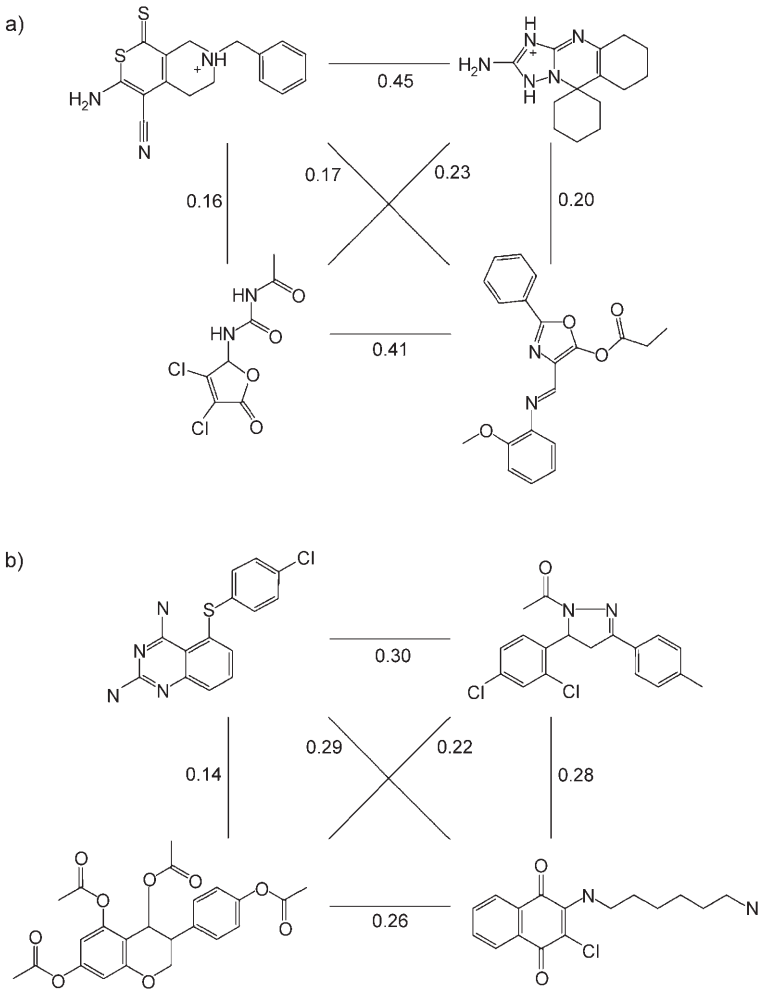

Figure 5. Structural diversity of HTS hits. Examples of hits are shown for the cathepsin B (top) and DHFR (bottom) screening data sets. Pair-wise MACCS Tanimoto similarities are reported.

to support similarity searching using fingerprints and aid in the identification of novel hits.

Keywords: fingerprints · molecular similarity · probabilistic modeling · screening data · virtual screening

[1] P. Willett, J. M. Barnard, G. M. Downs, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983 – 996.
[2] J. Bajorath, *Nat. Rev. Drug Discovery* **2002**, *1*, 882 – 894.
[3] P. Willett, *J. Med. Chem.* **2005**, *48*, 4183 – 4199.
[4] P. Willett, *Drug Discovery Today* **2006**, *11*, 1046 – 1053.
[5] J. Hert, P. Willet, D. J. Wilton, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177 – 1185.
[6] A. Schuffenhauer, P. Floersheim, P. Acklin, E. Jacoby, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391 – 405.
[7] L. Xue, J. W. Godden, F. L. Stahura, J. Bajorath, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1218 – 1225.
[8] H. Eckert, J. Bajorath, *Drug Discovery Today* **2007**, *12*, 225 – 233.
[9] J. M. Barnard, G. M. Downs, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 141 – 142.
[10] J. L. Durant, B. A. Leland, D. R. Henry, J. G. Nourse, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273 – 1280.
[11] MACCS structural keys, MDL Elsevier, San Leandro, CA, (USA), **2005**; http://www.mdl.com.
[12] C. A. James, D. Weininger, "Daylight Theory Manual, Version 4.9", Daylight Chemical Information Systems Inc., **2006**; http://www.daylight.com/dayhtml/doc/theory.
[13] L. Xue, J. W. Godden, F. L. Stahura, J. Bajorath, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1151 – 1157.
[14] A. Bender, Y. Mussa, R. C. Glen, S. Reiling, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708 – 1718.
[15] H. Eckert, J. Bajorath, *J. Chem. Inf. Model.* **2006**, *46*, 2515 – 2526.
[16] J. S. Mason, I. Morize, P. R. Menard, D. L. Cheney, C. Hulme, R. F. Labaudiniere, *J. Med. Chem.* **1999**, *42*, 3251 – 3264.
[17] E. K. Bradley, P. Beroza, J. E. Penzotti, P. D. J. Grootenhuis, D. C. Spellmeyer, J. L. Miller, *J. Med. Chem.* **2000**, *43*, 2770 – 2774.
[18] M. Baroni, G. Cruciani, S. Sciabola, F. Peruccio, J. S. Mason, *J. Chem. Inf. Model.* **2007**, *47*, 279 – 294.
[19] R. P. Sheridan, S. K. Kearsley, *Drug Discovery Today* **2002**, *7*, 903 – 911.
[20] J. W. Godden, J. Bajorath, *J. Chem. Inf. Model.* **2006**, *46*, 1094 – 1097.
[21] M. Vogt, J. W. Godden, J. Bajorath, *J. Chem. Inf. Model.* **2007**, *47*, 39 – 46.
[22] R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, 2nd ed., Wiley-Interscience, New York, **2000**.
[23] S. Kullback, *Information Theory and Statistics*, Dover Publications, Mineola, MN, **1997**.
[24] T. M. Cover, J. A. Thomas, *Elements of Information Theory*, Wiley-Interscience, New York, **1991**.
[25] M. Vogt, J. Bajorath, *J. Chem. Inf. Model.* **2007**, *47*, 337 – 341.
[26] A. Ormerod, P. Willett, D. Bawden, *Quant. Struct.-Act. Relat.* **1989**, *8*, 115 – 129.
[27] J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, A. Schuffenhauer, *J. Chem. Inf. Model.* **2006**, *46*, 462 – 470.
[28] A. Bender, H. Y. Mussa, R. C. Glen, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170 – 178.
[29] Molecular Operating Environment (MOE), Version 2005.06, Chemical Computing Group Inc., Montreal (Canada), **2005**; http://www.chemcomp.com.
[30] R. P. Sheridan, M. D. Miller, D. J. Underwood, S. K. Kearsley, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128 – 136.
[31] J. J. Irwin, B. K. Shoichet, *J. Chem. Inf. Model.* **2005**, *45*, 177 – 182.
[32] High-throughput screening assay for inhibitors of the thiol protease cathepsin B, University of Pennsylvania, **2007**; http://www.seas.upenn.edu/~pcmd/.
[33] High-throughput screening assay for inhibitors of the thiol protease cathepsin B, National Center for Biotechnology Information, Bethesda, MD (USA), **2007**; http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=453.
[34] M. Zolli-Juran, J. D. Cechetto, R. Hartlen, D. M. Daigle, E. D. Brown, *Bioorg. Med. Chem. Lett.* **2003**, *13*, 2493 – 2496.
[35] N. H. Elowe, J. E. Blanchard, J. D. Cechetto, E. D. Brown, *J. Biomol. Screening* **2005**, *10*, 653 – 657.
[36] J. W. Godden, F. L. Florence, J. Bajorath, *J. Chem. Inf. Model.* **2005**, *45*, 1812 – 1819.
[37] Molecular Drug Data Report (MDDR), MDL Elsevier, San Leandro, CA (USA), **2005**; http://www.mdl.com.
[38] L. Xue, J. Bajorath, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 757 – 764.
[39] Comprehensive Medicinal Chemistry Database (CMC-3D), Version 99.1, MDL Elsevier, San Leandro, CA (USA), **1999**; http://www.mdl.com.
[40] Ki database, National Institute of Mental Health, Psychoactive Drug Screening Program, Bethesda, MD, USA, **2006**; http://pdsp.med.unc.edu.
[41] X. Chen, M. Liu, M. K. Gilson, *Comb. Chem. High Throughput Screen.* **2001**, *4*, 719 – 725; http://www.bindingDB.org.
[42] Synthline Drug Database on STN International, taken from *Drugs of the Future* comprehensive drug monographs, Prous Science Barcelona, Spain, **1984**–present.

Received: April 19, 2007
Revised: May 10, 2007
Published online on June 11, 2007

