JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

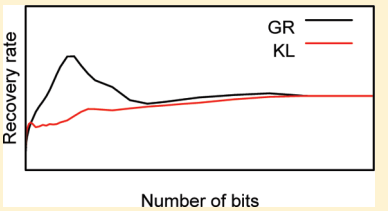
How Do 2D Fingerprints Detect Structurally Diverse Active Compounds? Revealing Compound Subset-Specific Fingerprint Features through Systematic Selection

Kathrin Heikamp and Jürgen Bajorath*

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

Supporting Information

ABSTRACT: In independent studies it has previously been demonstrated that two-dimensional (2D) fingerprints have scaffold hopping ability in virtual screening, although these descriptors primarily emphasize structural and/or topological resemblance of reference and database compounds. However, the mechanism by which such fingerprints enrich structurally diverse molecules in database selection sets is currently little understood. In order to address this question, similarity search calculations on 120 compound activity classes of varying structural diversity were carried out using atom environment fingerprints. Two feature selection methods, Kullback−Leibler divergence and gain ratio analysis, were applied to systematically reduce these fingerprints and generate alternative versions for searching. Gain ratio is a feature selection method from information theory that has thus far not been considered in fingerprint analysis. However, it is shown here to be an effective fingerprint feature selection approach. Following comparative feature selection and similarity searching, the compound recall characteristics of original and reduced fingerprint versions were analyzed in detail. Small sets of fingerprint features were found to distinguish subsets of active compounds from other database molecules. The compound recall of fingerprint similarity searching often resulted from a cumulative detection of distinct compound subsets by different fingerprint features, which provided a rationale for the scaffold hopping potential of these 2D fingerprints.



INTRODUCTION

Molecular fingerprints have for long been used for chemical similarity searching.[1,2] These descriptors typically consist of bit string representations of structural features or other molecular properties. Fingerprint representations calculated from molecular graphs, thus termed two-dimensional (2D) fingerprints, were among the early descriptors for similarity searching.[1] Original 2D fingerprint designs, such as structural keys,[3] were based on fragment dictionaries. In such fingerprints, each bit accounts for the presence or the absence of a predefined substructure in a compound. In addition to dictionary-based fingerprints, the introduction of topological 2D fingerprints that assemble connectivity pathways through molecules and represent them in a hashed format[4] has been another milestone event in this field. To this date, most—but not all—available 2D fingerprints account for structural and/or topological features.[2,5]

In similarity searching, the overlap between fingerprints of reference and database compounds is quantified as a measure of molecular similarity,[1] and database compounds are ranked in the order of decreasing fingerprint similarity to reference molecule(s) such that the structurally most similar compounds are highest on the list. As is generally the case with structural descriptors, 2D fingerprints do not capture biological activity information, and hence there is no well-defined relationship between (calculated) fingerprint similarity and (observed) biological activity similarity.[5,6] Of course, because fingerprints detect compounds that are structurally most similar to active reference molecules, these compounds have a certain probability to exhibit a similar activity. However, as structural similarity decreases between reference and ranked database compounds, calculated similarity and activity similarity are not related to each other, and it is generally difficult to select active compounds.[6]

Nevertheless, 2D fingerprints have a history of successful applications in virtual screening for novel active compounds.[2,5,6] Here the identification of structurally highly similar or analogous compounds, which one can easily accomplish using fingerprints, is much less interesting than the search for structurally diverse molecules having similar activity, a challenge often referred to as scaffold hopping.[7,8] However, although the scaffold hopping ability of relatively simple structural descriptors, such as 2D fingerprints, has often been questioned, it has clearly been demonstrated that 2D fingerprints are capable of enriching structurally diverse active compounds in small database selection sets, both in benchmark trials[9,10] and prospective virtual screening applications.[11,12] Methodological foundations for the rather surprising virtual screening potential of 2D fingerprints have, however, largely remained unclear.

The virtual screening performance of 2D fingerprints has been much improved over the years through the introduction of search

Received: June 16, 2011
Published: July 27, 2011

strategies for multiple reference compounds[13−15] and various new fingerprint designs.[6,16] Currently, fingerprints that capture atom environment information, such as Molprint2D[17,18] and especially extended connectivity fingerprints (ECFPs),[19] often produce the highest compound recall in comparative benchmark trials.[9,10,16] These types of fingerprints also capture topological information, similar to (yet algorithmically distinct from) the prototypic atom pathway fingerprints.[4] ECFPs systematically determine circular atom environments up to a given bond diameter in compounds and assemble these structural/topological features in a molecule-specific manner.[19] Feature arrays resulting from different compounds are then also quantitatively compared on the basis of Tanimoto similarity[1] or other standard similarity metrics.

Furthermore, the virtual screening performance of 2D fingerprints has also been increased through the introduction of fingerprint engineering strategies that modify fingerprint formats in specific ways, for example, by eliminating fingerprint bits (features) that are not critical for detecting a specific biological activity (fingerprint reduction)[20] or by combining the most important bit segments from fingerprints of different design (fingerprint hybridization).[21] Such modifications convert generally applicable fingerprints into compound class-specific versions with increased class-specific recall performance.[6,16] Such fingerprint engineering techniques have revealed that individual bit positions influence the outcome of similarity search calculations in different ways, depending on the compound classes under investigation.[6,16]

Fingerprint reduction techniques depend on the application of feature ranking and selection methods that make it possible to evaluate the importance of individual bits (features) for detecting compounds belonging to a given activity class. For fingerprint reduction, Kullback−Leibler (KL) divergence analysis[22] from information theory[23] has been originally applied[20] and has thus far been a method of choice.[16]

In this study, we have carried out comparative feature selection analysis for atom environment fingerprints and a large number of activity classes to revisit the scaffold hopping potential of 2D fingerprints and address the question why such fingerprints are capable of recognizing active compounds having little structural and topological resemblance to reference molecules. We have compared KL divergence with gain ratio (GR)[24] analysis, another information−theoretic approach, for activity classes of varying degrees of structural diversity and monitored compound recall characteristics of systematically reduced fingerprints. On the basis of this analysis, we have found that for structurally diverse activity classes, small numbers of fingerprint features are responsible for distinguishing different subsets of active compounds from the background database, resulting in a cumulative detection of such compound subsets. Taken together, these findings suggest a plausible mechanism for scaffold hopping by state-of-the-art 2D fingerprints.

METHODS AND MATERIALS

Fingerprints. For our analysis, two atom environment fingerprints were selected, Molprint2D[17,18] and an ECFP with bond diameter four (ECFP4).[19] These 2D fingerprints have often yielded high compound recall rates in comparative fingerprint benchmark investigations and are considered state-of-the-art.[16] Molprint2D was calculated using public domain software tools[17] and ECFP4 using Pipeline Pilot.[26]

Activity Classes. A total of 120 activity classes, each containing at least 200 compounds active against human targets, was extracted from BindingDB.[25] Selected compounds were required to be rule-of-five compliant, have at least 1 μM potency (K_i or IC_{50} values), and consist of atom types compatible with the calculation of the Molprint2D fingerprint.[17] Although fingerprint search calculations do not take potency information into account, the potency threshold was applied to avoid the inclusion of very weakly or borderline active compounds in similarity searching. Atom typing and rule-of-five calculations were carried out using Pipeline Pilot.

The relative structural diversity of activity classes was assessed by determining the number of Bemis−Murcko scaffolds (BMS)[27] and corresponding carbon skeletons (CSK), also referred to as cyclic skeletons,[27,28] per class and by calculating the average ratio of compounds per BMS and CSK. BMS consist of all rings and linker fragments between rings after removal of R-groups from compounds. They are further reduced to CSK by setting all bond orders to one and by converting all heteroatoms to carbon.

Feature Ranking Methods. For fingerprint feature ranking, KL divergence[22] and GR analysis[24] were carried out. Both methods determine the ability of individual fingerprint bits/features to differentiate between active and background database (inactive) compounds. In the following, the terms bit and feature are synonymously used.

The KL divergence measures the difference between two value distributions $p(x)$ and $q(x)$. When $p(x)$ describes the probability of a value (0 or 1) of bit x in active and $q(x)$ the probability of a value of this bit in inactive compounds, the KL divergence D is defined as

$$D(p(x)||q(x)) = \sum_x p(x) \log_2 \frac{p(x)}{q(x)}$$

Hence, bit positions can be ranked on the basis of D to assign high priority to bits that are preferentially set on in active compounds. KL divergence calculations have thus far been applied for the identification of bit settings that are characteristic of activity classes and fingerprint reduction.[20,21]

Furthermore, GR is a statistical feature ranking approach that is based on normalized mutual information[23] (MI). MI measures the amount of information a variable X contributes to the values of another variable Y. Given the definition that X describes the value of a fingerprint bit (0 or 1) and Y the activity states (active or inactive), MI determines how much information about the correct activity state of test compounds is specified by a bit:

$$\begin{aligned} MI(X;Y) &= H(Y) - H(Y|X) \\ &= \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \end{aligned}$$

where H is the information entropy, $p(x)$ and $p(y)$ are the probability functions of X and Y, and $p(x,y)$ the joint probability function. Thus, MI represents the difference between the entropy of the activity states and the entropy of the activity states under the condition that the value of a specific bit is known. GR is then obtained by dividing MI by the overall entropy of this bit:

$$\begin{aligned} GR(X;Y) &= \frac{MI(X;Y)}{H(X)} \\ &= \frac{-\sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}}{\sum_x p(x) \log_2 p(x)} \end{aligned}$$

For a given activity class, bits are ranked on the basis of their KL or GR values, which reflect the ability of each individual bit to distinguish between active and database compounds. For example, if a bit would occur in all active but no database compounds, then this ability would be maximal.

Reduced Fingerprints. For each activity class, reduced Molprint2D and ECFP4 fingerprints were generated on the basis of the KL and GR feature rankings by incrementally extending sets of highly ranked bits. Bit probability distributions within active and database compounds were derived on the basis of the active reference set and the database (without active compounds), respectively. In order to adjust probabilities of zero, which would lead to nondefined KL divergence and GR calculations, an *m*-estimate correction was applied by adding a (hypothetical) fingerprint reflecting the average fingerprint bit settings of the database compounds to the fingerprints of each reference set (and an equivalent correction was applied to the database). In addition, prior to feature ranking, bits were removed that only occurred in very few databases but no active compounds because their calculated feature weights were negligible.

Following feature ranking, reduced fingerprints were generated as follows, beginning with the smallest possible versions: The top five bits were added one-by-one, hence generating five minimal fingerprint versions consisting of only one to five bits. Then, bits from rank 6 to 20 were added in 5-bit increments, hence generating three further extended fingerprint versions. Furthermore, bit positions ranked from 21 to 100 were added in 10-bit increments, bits ranked from 101 to 200 in 20-bit, bits ranked from 201 to 500 in 50-bit, and bit positions ranked from 501 to 1000 in 100-bit increments. Thus, in the design of reduced fingerprints, the most highly ranked bits were utilized on an individual basis, and bit positions of decreasing significance were considered in increments of increasing size. For each activity class and original fingerprint, 32 reduced fingerprint versions were generated. Thus, in total, 7680 reduced fingerprints were investigated, in addition to the unmodified Molprint2D and ECFP4 fingerprints.

Fingerprint Similarity Searching. From each activity class, 100 different subsets of 20 compounds each were randomly selected as reference sets for 100 independent similarity search trials. In each case, the remaining active compounds were added as potential hits to a background database consisting of 1.44 million compounds randomly selected from ZINC.[29] As a search strategy, 10 nearest neighbor (10NN) calculations[13,14] were carried out. Following this strategy, each database compound is compared to each individual reference molecule by calculating pairwise Tanimoto similarity,[1] and the final similarity score of a database compound is obtained by averaging the 10 highest individual similarity values. Compound recall rates were determined for a selection set of 5000 database compounds (i.e., ~0.35% of the background database). This selection set size was chosen because several of the activity classes contained more than 1000 compounds. Because of the differences in activity class size, there were different probabilities for the random enrichment of active compounds in database selection sets. However, since we did not compare search results across different activity classes but analyzed the performance of unmodified and reduced fingerprints on individual activity classes, differences in random enrichment probabilities across different classes did not need to be considered in the context of our analysis. Importantly, for ranking of active compounds, we applied a pessimistic ranking strategy such that all background database compounds were included in the ranking prior to the last recovered active compound within a selection set. This means that compounds having the same similarity value were not given the same but subsequent ranks. Otherwise, the top 5000 ranks might yield more than 5000 compounds, which would bias the search results.

Given the number of different fingerprint versions and reference sets for each activity class, a total of 792 000 similarity search trials were carried out for our analysis.

RESULTS AND DISCUSSION

Study Objective. We have been interested in exploring the question of how compound recall characteristics of fingerprint search calculations might be rationalized, with a particular focus on scaffold hopping ability. In order to generate a substantial body of primary similarity search data for further analysis, we have initially carried out systematic fingerprint search calculations on 120 different activity classes using 2 state-of-the-art 2D fingerprints. To obtain statistically meaningful results, 100 different reference sets per activity class were utilized. On the basis of these data, compound recall characteristics have then been explored in detail. We have systematically applied and compared feature selection methods to identify fingerprint features that were responsible for the recall of different compound classes. In this context, we have then focused on the question why 2D fingerprints that emphasize structural/topological resemblance have the potential to enrich structural diverse active compounds in database selection sets. In the following, we describe the composition of data sets used for our analysis, present the results of large-scale fingerprint searching, discuss the findings of comparative feature selection, compare the search performance of unmodified and reduced fingerprint representations, and attempt to rationalize the scaffold hopping potential of the 2D fingerprints studied here.

Composition of Data Sets. The composition of the 120 activity classes utilized for systematic fingerprint similarity searching is reported in Table S1 of the Supporting Information. In the text, activity classes are referred to by numbers given in Table S1, Supporting Information. These compound data sets contained between 200 and 1425 different enzyme inhibitors or receptor ligands with varying degrees of intraclass structural diversity, reflected by differences in BMS and CSK distributions and compound-to-BMS and -CSK ratios, as also reported in Table S1, Supporting Information. For our analysis, we distinguish between structurally more homogeneous and heterogeneous (diverse) compound classes. Therefore, compound-to-scaffold ratios were determined to estimate intraclass structural diversity. This approach was considered more robust than, for example, the calculation of average compound similarity values. For example, if a data set would consist of a few topologically distinct scaffolds, each of which would be represented by a larger number of analogs, then average pairwise similarity values might be relatively low, although intraclass structural diversity would be limited in this case. However, the compound-to-scaffold ratio would be rather high, which would better account for limited intraclass diversity. In our set of activity classes reported in Table S1, Supporting Information, structural homogeneous classes are characterized by the presence of comparably small numbers of BMS and CSK and large compound-to-BMS and -CSK ratios, with about 5−10 or more compounds per CSK. For activity classes of increasing structural diversity, the compound-to-CSK ratio is decreasing to about 2−3 compounds per CSK.

Table 1. Average Recovery Rates for ECFP4[a]

	unmodified	GR		KL	
no.	RR	RR	no. bits	RR	no. bits
1	45.5	69.3	120	45.3	800
2	50.8	75.4	140	53.8	700
3	37.6	61.1	140	37.6	700
4	49.4	65.0	140	46.9	700
5	96.6	98.7	600	98.5	60
6	51.8	67.2	120	50.6	800
7	52.4	67.7	120	50.3	800
8	81.9	84.1	90	79.9	700
9	72.1	84.6	80	71.4	350
10	60.1	60.5	700	59.6	700
11	74.2	74.0	700	73.6	700
12	51.6	66.0	120	50.8	700
13	71.7	76.6	160	69.7	800
14	93.8	95.8	120	94.4	700
15	68.0	67.0	140	65.6	700
16	58.4	77.7	120	64.0	140
17	86.9	87.5	600	87.9	700
18	93.3	97.3	80	95.8	40
19	48.1	61.0	140	45.1	900
20	58.2	66.1	120	55.4	800
21	58.2	62.4	700	60.9	700
22	49.5	55.6	700	53.7	700
23	65.9	70.0	600	67.9	700
24	79.6	92.8	80	80.2	60
25	64.2	87.9	140	67.1	700
26	62.5	88.5	120	70.7	120
27	73.6	91.1	120	81.9	80
28	67.6	80.3	120	68.1	700
29	72.4	80.4	100	73.2	700
30	85.2	85.9	140	85.0	600
31	52.5	65.6	140	54.7	800
32	44.0	64.8	140	50.3	800
33	79.1	92.6	120	82.3	90
34	76.3	76.3	700	76.7	700
35	72.9	74.7	800	74.3	800
36	87.7	93.1	70	88.0	350
37	71.6	92.0	160	78.5	40
38	78.2	82.2	100	78.7	600
39	38.6	42.5	120	40.0	700
40	34.6	35.7	700	36.3	700
41	30.0	34.4	120	34.0	700
42	53.0	64.7	120	49.2	700
43	71.3	72.1	140	71.0	600
44	85.1	85.1	90	83.0	800
45	69.6	71.2	60	65.1	600
46	79.5	78.4	800	78.1	700
47	79.3	78.0	800	77.6	700
48	55.6	77.8	120	65.5	700
49	77.0	83.8	120	76.2	700
50	53.2	80.7	200	58.0	800
51	85.4	85.4	90	84.0	600
52	63.8	75.1	120	62.2	600
53	87.0	92.2	160	87.0	700
54	27.7	39.3	160	29.4	800
55	48.6	63.0	100	48.5	700
56	64.7	93.4	70	85.2	15
57	75.0	81.9	100	72.7	700
58	82.2	92.0	90	80.2	700
59	85.6	89.9	600	90.0	600
60	88.2	90.8	250	89.9	600
61	81.2	84.1	90	82.5	80
62	64.8	77.1	180	68.3	800
63	61.3	81.7	120	60.3	140
64	82.9	84.5	600	84.9	600
65	36.9	54.7	120	43.2	600
66	64.2	93.8	140	85.5	10
67	67.9	83.5	140	71.4	700
68	53.8	82.3	160	67.4	3
69	73.7	90.2	100	78.2	10
70	56.3	86.5	140	76.8	1
71	39.4	56.0	120	41.4	800
72	78.5	81.0	90	74.0	800
73	71.6	75.2	90	66.4	800
74	66.2	70.6	90	62.7	700
75	78.4	93.5	60	82.8	40
76	66.9	82.3	90	69.1	350
77	66.3	71.3	180	69.3	800
78	45.7	64.7	140	44.7	800
79	50.6	67.6	160	48.4	800
80	69.7	79.3	120	67.8	800
81	84.5	93.9	120	85.4	50
82	49.6	64.4	120	46.0	700
83	55.7	82.4	100	58.0	600
84	72.8	80.5	140	72.4	800
85	61.4	79.1	140	61.0	800
86	50.9	70.1	160	48.0	700
87	62.5	76.2	70	57.5	700
88	42.7	68.8	120	43.4	50
89	80.1	88.0	60	80.5	700
90	63.0	71.1	120	61.2	800
91	67.0	82.0	140	65.3	700
92	64.1	82.9	100	64.9	700
93	69.2	76.0	160	67.5	800
94	71.3	82.1	160	75.5	800
95	77.0	88.8	80	77.0	30
96	94.1	95.4	60	93.5	140
97	37.1	48.0	140	39.0	800
98	75.9	76.0	80	75.5	700
99	76.8	80.4	90	76.6	700
100	28.9	45.9	120	30.5	800
101	52.8	68.1	120	46.3	700
102	33.8	53.5	120	34.7	700
103	65.0	70.0	120	66.7	700
104	41.1	44.4	100	40.1	800
105	39.2	61.7	140	38.8	800

Table 1. Continued

	unmodified	GR		KL	
no.	RR	RR	no. bits	RR	no. bits
106	35.6	40.2	120	33.9	800
107	85.9	87.2	80	85.0	600
108	41.9	73.1	160	54.0	160
109	85.2	87.0	700	86.5	700
110	77.3	84.4	100	77.5	700
111	80.3	90.4	100	87.3	600
112	78.5	97.5	120	91.5	20
113	55.2	68.6	120	56.3	200
114	58.5	68.0	140	59.2	800
115	64.8	73.6	80	63.7	700
116	73.1	72.5	450	72.4	600
117	81.0	91.7	80	89.6	40
118	64.7	62.9	500	64.0	700
119	59.4	67.3	140	58.8	700
120	27.8	45.6	140	29.8	450

[a] Average recovery rates (RR) over 100 independent trials are reported for full-length (unmodified) ECFP4 and the best-performing reduced fingerprints derived by GR and KL divergence. The length of each reduced fingerprint (no. bits) is specified. Recovery rates are calculated for a database selection size of 5000 compounds.

Large-Scale Similarity Searching. We first compared the overall search performance of unmodified ECFP4 and Molprint2D. Average recall rates over 120 activity classes were 63.0% and 64.5% for Molprint2D and ECFP4, respectively. ECFP4 produced higher recall rates for 70 classes and Molprint2D for 47 classes (with equal performance in three cases). Hence, both fingerprints reached comparable performance levels, but ECFP4 produced overall slightly better results. All search results for ECFP4 and Molprint2D are provided in Table 1 and Table S2 of the Supporting Information, respectively. In addition, histogram representations of recall rates according to Table 1 are provided in Figure 1a and b for 10 activity classes with the highest and lowest increase in recall, respectively, as a consequence of GR feature selection.

Global Effects of Feature Selection. Both Molprint2D and ECFP4 are combinatorial fingerprints that can, in principle, yield exceedingly large feature numbers (although this is typically not the case for small organic compounds). For feature selection studies, we considered the top 1000 features on the basis of GR and KL divergence ranking. We generally observed that reduced fingerprints, often of only small size, met or exceeded the search performance of unmodified ECFP4 and Molprint2D, as reported in Table 1 and Table S2 of the Supporting Information. However, there were notable global differences between the two alternative feature selection approaches. On average, the best-performing reduced fingerprints selected on the basis of KL divergence consisted of 252 and 578 bits for Molprint2D and ECFP4, respectively. For GR selection, the corresponding numbers were 182 bits for Molprint2D and 196 bits for ECFP4. Thus, the best-performing reduced fingerprints generated on the basis of GR contained fewer features than those generated on the basis of KL divergence. Furthermore, for Molprint2D and ECFP4, the best KL divergence-based fingerprints produced average recall rates of 65.3% and 65.9%, respectively, which slightly increased the recall rates of the full-length fingerprints (63.0% and 64.5%, respectively). However, GR-based reduced fingerprints achieved average recall rates of 74.3% and 75.4% for Molprint2D and ECFP4, respectively. Thus, the top-performing GR-based fingerprints consisted of fewer than 200 features and further increased the average recall rates of the unmodified fingerprints by approximately 10%. Given these differences observed in feature selection, we compared KL divergence and GR approaches in more detail.

Comparison of Feature Probabilities. Next we studied the probabilities of top-ranked bits selected by KL divergence and GR to occur in active and database compounds. For this purpose, many individual search trials on our activity classes were analyzed (averages over different references sets leading to different bit selections would not be meaningful to calculate). In Figure 2, the corresponding probabilities of the 100 top-ranked KL divergence and GR bits of ECFP4 are compared for two different activity classes. These results are representative of many comparisons we carried out and reflect a clear general trend we observed. It should be noted that the probabilities in Figure 2 are reported on a logarithmic scale. Because of the large number of database compounds, the probabilities of bit settings in the database might become very small. In Figure 2a, the probability histogram for the GR selection of ECFP4 features for a reference set of activity class no. 24 is shown, and in Figure 2b, the corresponding histogram for the KL divergence selection is shown. In Figure 2a, the top 25 bits in the GR-based histogram have a decreasing probability of occurrence in active compounds but no probability of occurrence in database compounds. Over the remaining bit positions, the active probability remains high, and the database probability slightly increases. Thus, the GR-based bit ranking reflects strong emphasis on probability differences between active and database compounds. By contrast, the corresponding KL divergence-based bit ranking in Figure 2b reveals a less systematic profile. In this case, bit positions are also highly ranked that have a detectable probability to occur in database compounds. In these cases, however, the active probabilities are very high. Other bit positions with lower active probability but no database probability occur at intermediate ranks. This profile phenotype can be rationalized on the basis of the KL divergence formula presented in the Methods and Materials Section. Thus, KL divergence calculations not only emphasize probability differences but also the magnitude of the active probability, ultimately leading to a selection compromise. Equivalent observations concerning the GR- and KL divergence-based probability histograms are made in Figure 2c and d, respectively, for a reference set of activity class 22 (that is structurally more homogeneous than class 24).

Hence, there were significant differences between the GR and KL divergence feature selections, which also applied to the actual features that were prioritized. We generally observed that the overlap between GR- and KL divergence-based rankings considerably varied for different activity classes. This is illustrated in Table 2 where the average GR versus KL divergence overlap of features selected for the 100 individual reference sets of activity classes 22 and 24 is reported. For the 20 most significant bits, the overlap was only small for class 22 but large for class 24. Taking bits 30−100 into account, the overlap is ultimately increasing to ∼89% and ∼84%. Thus, most significant differences were observed for top-ranked bit positions.

Considering the differences between the GR- and KL divergence-based histograms in Figure 2, we conclude that GR provides a more stable feature selection approach for fingerprint reduction. This is the case because GR yields bit rankings that

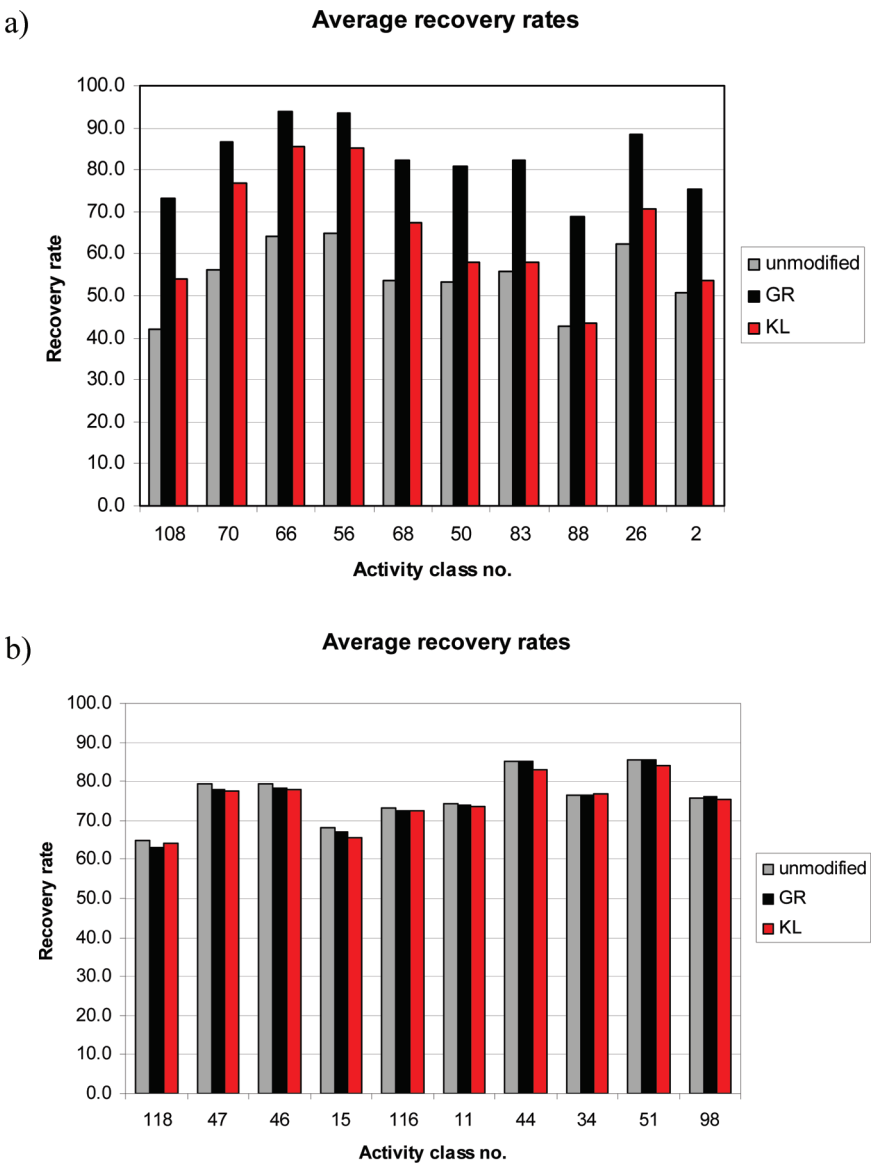


Figure 1. Average recovery rates for ECFP4. Shown are 10 activity classes with the highest (a) and lowest (b) increase in recovery rates as a consequence of GR feature selection. Also shown are the corresponding recovery rates for unmodified fingerprints and KL-based selection.

predominantly reflect probability differences between active and database compounds and are more intuitive than KL divergence rankings. This conclusion is also consistent with the overall better search performance we observed for fingerprints reduced on the basis of GR compared to KL divergence and their smaller size, as discussed above. However, on the basis of KL divergence analysis, reduced fingerprints, in part, with significantly increased search performance compared to the original fingerprints have also been generated in a number of cases,[20,21] thus demonstrating its feature selection potential. In addition, KL divergence analysis is generally less dependent on background database composition than GR because it emphasizes the magnitude of active probabilities, which should also be taken into account in cases where background databases are relatively small.

Compound Recall Characteristics. We next analyzed recovery rates for reduced fingerprints consisting of up to 1000 bit positions. For each activity class, averages were calculated for all 100 reference sets. Representative examples are shown in Figure 3. We generally observed that recovery rates peaked at relatively small feature numbers and then decreased and/or remained constant as feature numbers increased. The recovery rate profiles were overall comparable for both ECFP4 and Molprint2D, although details differed in many cases. For both fingerprints, a notable difference between GR and KL divergence selection was also observed at the level of recall curves. GR-based feature selection often led to a sharper increase in recall performance, resulting in a clear peak followed by a reduction in recovery rates for further increasing feature numbers. Then the recovery rates essentially remained constant. These recall characteristics were frequently also observed for KL divergence selection but were generally less obvious. Figure 3a−h illustrates these effects. The number of features required to reach the top search performance was often found to differ between structurally diverse and homogeneous activity classes. In Figure 3a, ECFP4 search results are shown for activity class 1 (with a compound-to-CSK ratio of 3.36), and in Figure 3c, ECFP4

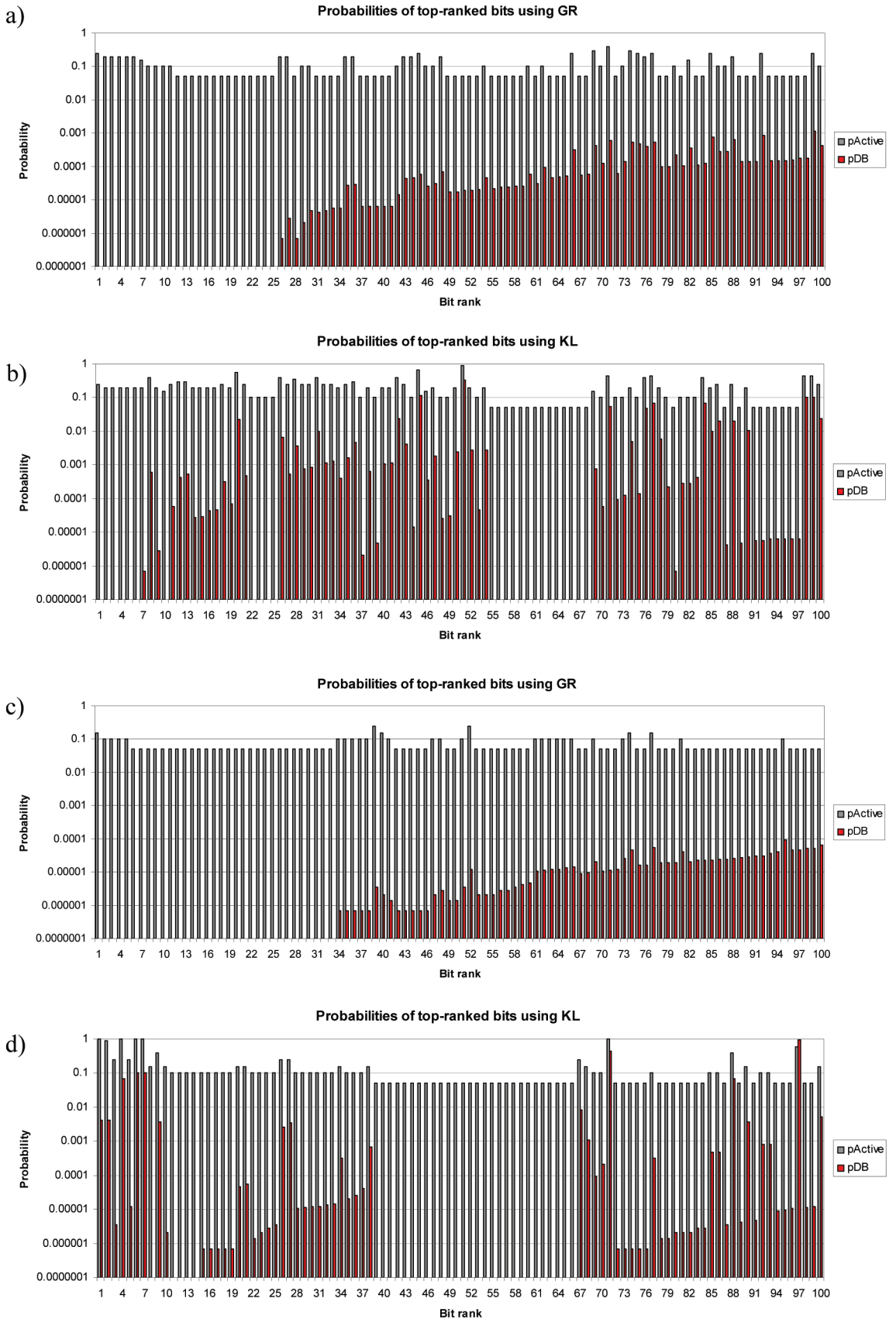


Figure 2. Probabilities of top-ranked bits. For two exemplary activity classes of different structural diversity (no. 24, diverse; no. 22, homogeneous) and an individual reference set, the probabilities of the 100 top-ranked bits to occur in active (pActive) and database compounds (pDB) are reported for different feature selection methods. Probabilities (without *m*-estimate correction) are plotted on a logarithmic scale. (a) 24/GR, (b) 24/KL divergence, (c) 22/GR, (d) 22/KL divergence.

dx.doi.org/10.1021/ci200275m |*J. Chem. Inf. Model.* 2011, 51, 2254–2265

Table 2. Overlap of Bits Between GR and KL Divergence[a]

	overlap	
no. bits	no. 22	no. 24
1	0.0	100.0
2	2.5	90.0
3	6.7	84.0
4	7.5	79.3
5	7.2	75.4
10	20.8	61.4
15	27.7	51.8
20	35.5	49.8
30	53.3	53.3
40	70.9	58.2
50	76.8	60.6
60	80.8	70.4
70	82.7	75.9
80	86.7	78.3
90	88.3	80.5
100	88.8	84.1

[a] For two exemplary activity classes of different structural diversity (24, diverse; 22, homogeneous), the average overlap (in %) between features in reduced fingerprints of different size (no. bits) selected by GR and KL divergence is reported. The average overlap was calculated for 100 independent search trials with different reference sets, and in each case, the 100 top-ranked bit positions were compared.

results are shown for activity class 54 (ratio 2.13). Here GR-based reduced ECFP4 representations consisting of 120 bits for class 1 and 160 bits for class 54 reached a clear performance peak. In Figure 3e and f, corresponding ECFP4 search profiles are shown for activity class 30 (ratio 10.33) and class 75 (ratio 7.17), respectively. These structurally more homogeneous activity classes yielded higher compound recovery rates than the more diverse classes 1 and 54; class 30 required 140 ECFP4 features to reach the top performance and class 75 only 60 features.

Taken together, despite the fingerprint, selection method, and activity class dependent differences we observed, three general conclusions could be drawn for both fingerprints and GR selection from the findings discussed above. First, reduced fingerprints, rather than unmodified versions, generally produced highest recall rates. Second, structurally diverse activity classes often—but not always—required more features to reach top performance than structurally more homogeneous classes. Third, in all instances, the recall curves revealed a nearly linear increase in recovery rates over increasing bit numbers until the top search performance was reached. Depending on the activity classes, the slope of these pseudolinear curve intervals often differed, as illustrated in Figure 3.

Recovery of Activity Class Subsets. In order to further rationalize the observations discussed above, we analyzed the number of active database compounds (correctly identified hits) and other database molecules that were detected by reduced fingerprints of increasing size of up to 100 bits, beginning with the smallest representations containing the most highly ranked features. In Table 3, three representative examples are shown for ECFP4 features and GR selection. In Table 3a, activity class 1 (compound-to-CSK ratio 3.36) contained 554 compounds and yielded a recovery rate of 45.5% with full-length ECFP4 and 69.3% with the best reduced fingerprint. In Table 3b, class 24 (ratio 3.82) consisted of 294 compounds and produced an ECFP4 recovery rate of 79.6% and of 92.8% for the best reduced version. In Table 3c, class 36 (ratio 7.06) contained 374 compounds and yielded a recovery rate of 87.7% for ECFP4 and 93.1% for the best reduced fingerprint. Hence, these activity classes had different composition and displayed partly different recall characteristics. As reported in Table 3a−c, a varying number of the most highly ranked fingerprint features consistently detected a significant number of the active compounds, without selecting any other database molecules. For activity class 1, 24, and 36, the top 40, 20, and 10 bits exclusively recognized 206, 211, and 104 active compounds, respectively. Hence, the exclusive recognition of active compounds by small feature sets significantly contributed to the overall compound recall. As also revealed in Table 3, the inclusion of additional bits resulted in further detection of active compounds accompanied by a steady and, in part, dramatic increase in the number of other database molecules that were detected, corresponding to a substantial loss of the specificity of the search calculations. From these observations, one can infer that the generally observed gain in search performance through fingerprint reduction can largely be attributed to the elimination of fingerprint features that predominantly increase the background noise of the calculations (i.e., preferentially detect other database compounds).

Moreover, Table 3 reveals another trend that is highly relevant for our analysis. The most important fingerprint features were specific for subsets of active compounds. For example, in Table 3a, the first bit detected 45 active compounds, the second 80 other compounds, the third another 38 previously unrecognized actives, and so on. There was a steady increase in active compounds up to a size of 100 bits. Starting with 50 bits, database molecules were beginning to be retrieved. In the presence of 100 bits, the number of detected active compounds approximately doubled compared to 40 bits (when still no database compounds were detected), but 2090 other database molecules were also selected. In Table 3b, bit 1 detected 80 compounds, bit 2 selected 40 more, bit 3 did not add more active compounds, but bit 4 detected 37 additional ones, which then remained essentially unchanged for up to 15 bits, until the inclusion of the next 5 bits led to the detection of another 53 active compounds, still without retrieval of other database molecules. Beginning with 30 bits, other database molecules were detected, and the increments of newly recognized active compound became smaller. Corresponding effects also occurred for the structurally more homogeneous activity class in Table 3c, although there was overall less variation in the number of active compounds detected by increasing numbers of bits, with the exception of notable increases in active compounds within the range of 10−30 bits, where other database compounds were also selected.

Figure 4 illustrates some of the results discussed above. Shown are five structurally diverse hits that were recognized by the top three fingerprint bits according to Table 3a. Each of these compounds contains only one of the atom environments encoded by these three bits, i.e., it responds to only one of the three top-ranked bits. The corresponding substructures are mapped on the hits. The comparison reveals that a substructure match provided by a single fingerprint feature has been sufficient in these instances to facilitate a scaffold hop. With only the first three fingerprint bits, a total of 163 active compounds were retrieved that yielded 35 distinct CSKs.

Taken together, on the basis of GR selection, compound subset detection by bit subsets was consistently observed for

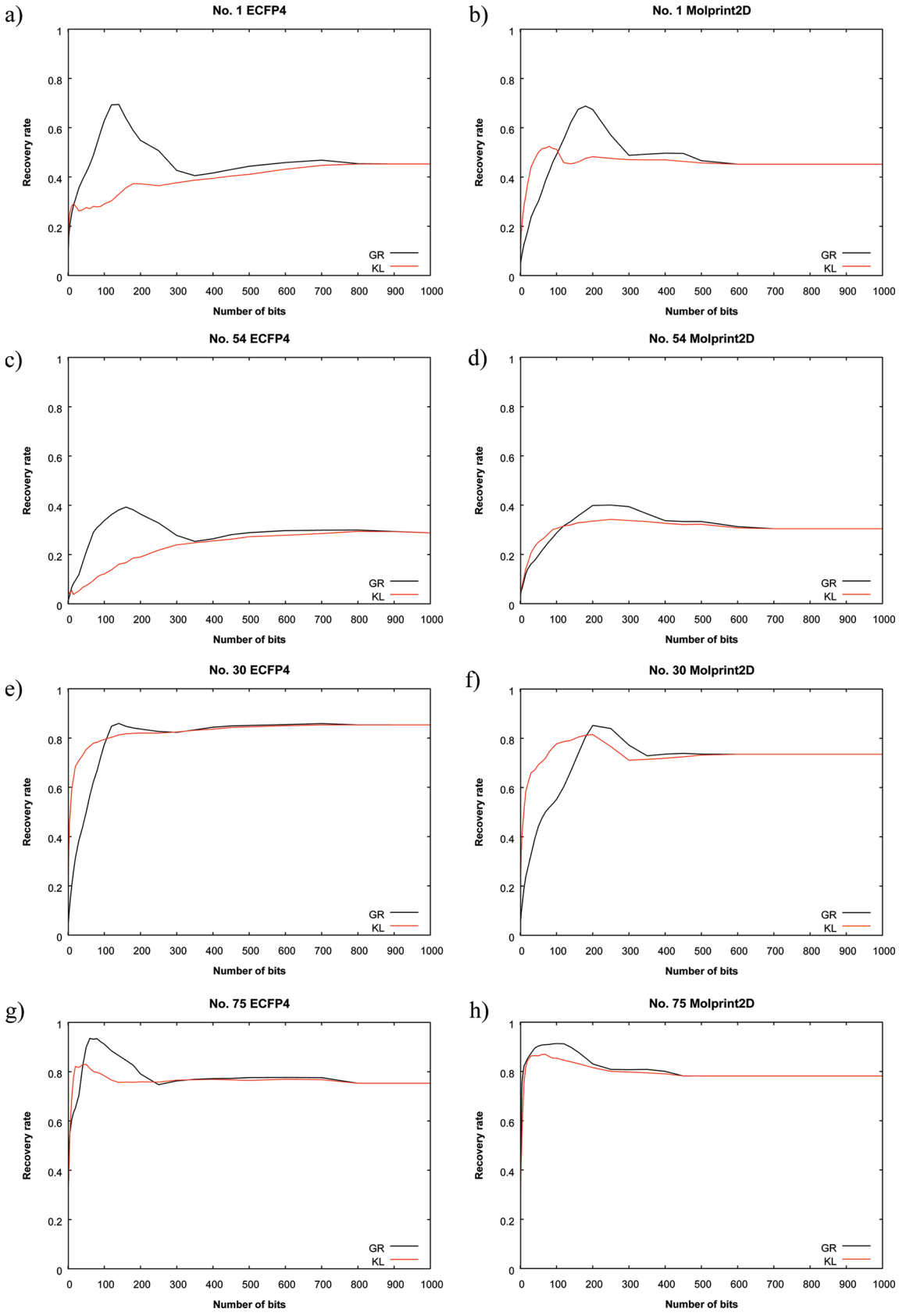


Figure 3. Recovery rates for reduced fingerprints. For four exemplary activity classes of different structural diversity (no. 1 and 54, diverse; no. 30 and 75, homogeneous), average recovery rates are reported for reduced fingerprint representations with increasing number of bits selected by GR or KL divergence. (a) 1/ECFP4, (b) 1/Molprint2D, (c) 54/ECFP4, (d) 54/Molprint2D, (e) 30/ECFP4, (f) 30/Molprint2D, (g) 75/ECFP4, and (h) 75/Molprint2D.

Table 3. Detection of Activity Class Subsets[a]

	GR	
no. bits	no. ADC	no. DC
(a) Activity Class 1		
1	45	0
2	125	0
3	163	0
4	176	0
5	176	0
10	176	0
15	184	0
20	187	0
30	206	0
40	206	0
50	212	9
60	243	38
70	293	132
80	308	379
90	352	860
100	425	2090
(b) Activity Class 24		
1	80	0
2	120	0
3	120	0
4	157	0
5	157	0
10	157	0
15	158	0
20	211	0
30	226	15
40	228	151
50	241	589
60	261	790
70	268	1936
80	271	4557
90	256	4503
100	243	4017
(c) Activity Class 36		
1	86	0
2	87	0
3	87	0
4	90	0
5	90	0
10	104	0
15	110	3
20	158	52
30	319	177
40	319	236
50	319	383
60	319	1648
70	319	4240
80	306	109
90	306	104
100	306	104

[a] For representative activity classes, the numbers of detected active database compounds (no. ADC) and other database compounds (no. DC) retrieved prior to the last recovered active are reported for varying numbers of ECFP4 features (no. bits) selected by GR. In each case, the results are shown for an individual reference set: (a) 1, diverse; (b) 24, diverse; and (c) 36, homogeneous.

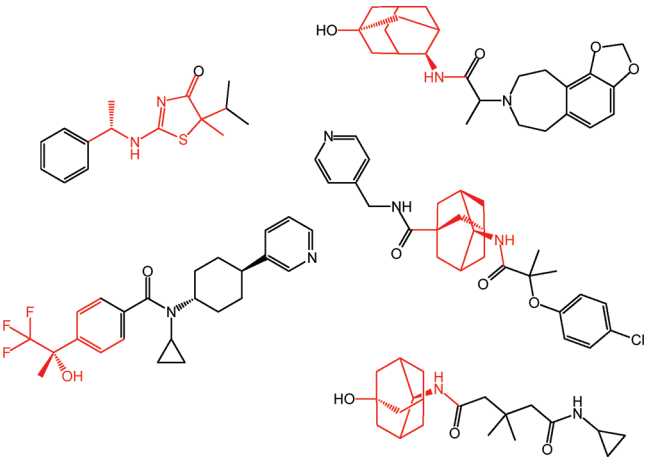


Figure 4. Structurally diverse hits. Shown are five exemplary hits detected with the top three fingerprint bits selected by GR, as reported in Table 3a, that represent scaffold hops. In the fingerprint of each of these compounds, only one of the three bits was set on, and the corresponding structural features are mapped on the hits (red).

reduced ECFP4 and Molprint2D fingerprints. On the basis of KL divergence selection, similar subsets effects were also found, in particular for structurally diverse activity classes but much less so for structurally homogeneous classes. Representative examples for KL divergence selection are provided in Table S3 of the Supporting Information. For structurally homogeneous activity classes, KL divergence selection often yielded top-ranked bits that were not specific for active compounds but also detected other database molecules, different from many structurally diverse classes. By contrast, GR always yielded bit positions that were specific for active compounds.

Scaffold Hopping Potential Revisited. The incremental recognition of different subsets of active compounds by small fingerprint feature sets observed in our analysis, especially for structurally diverse activity classes, provided a rationale for the scaffold hopping potential of the investigated fingerprints. On the basis of our findings, the overall recovery rates of active compounds achieved by these 2D fingerprints largely resulted from the cumulative detection of distinct subsets of active compounds by different fingerprint features. Typically, small numbers of key features specifically selected active compounds over other database molecules. Often, single bit positions were responsible for the detection of relatively large compound subsets. Other less specific bits also retrieved additional subsets of active compounds and also rapidly increased the number of other database molecules. Thus, identifying those bit positions

that were most important for recognizing different activity classes and analyzing their contribution to the recovery of active compounds revealed a plausible mechanism for scaffold hopping using the 2D fingerprints studied here.

CONCLUDING REMARKS

In this study, we have investigated in detail the compound recall characteristics of state-of-the-art 2D fingerprints. Beginning with a large-scale fingerprint search campaign, feature selection methods were applied to systematically reduce original fingerprint and identify the most important fingerprints bits/features for each activity class. Fingerprint reduction generally improved compound recovery rates, consistent with earlier findings. In many instances, small numbers of bits were sufficient to yield the highest search performance. In addition, our results indicated that fingerprint reduction mostly improved compound recall by omitting features that predominantly recognized other database compounds (and thus reduced the specificity of the search calculations). By comparing GR- and KL divergence-based fingerprint feature selection, we identified different characteristics of these approaches, assigning overall higher confidence to GR-based bit rankings. On the basis of feature selection, we observed, in particular, for structurally diverse activity classes, that small numbers of highly ranked fingerprint features (often individual bits) distinguished subsets of active compounds from other database molecules. Additional features also recognized distinct compound subsets but were not specific for active compounds. These cumulative subset contributions to compound recovery rationalized the scaffold hopping potential of 2D atom environment fingerprints and revealed a mechanism for the recognition of structurally diverse hits. It is anticipated that the feature selection approaches introduced herein will be useful for additional mechanistic studies on fingerprints of different design and for further fingerprint engineering applications.

ASSOCIATED CONTENT

Supporting Information. Tables S1, S2, and S3 report the composition of compound activity classes, average recovery rates for Molprint2D, and activity class subset detection on the basis of KL divergence, respectively. This information is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: bajorath@bit.uni-bonn.de. Telephone: +49-228-2699-306.

ACKNOWLEDGMENT

The authors thank Britta Nisius and Martin Vogt for many helpful discussions.

REFERENCES

(1) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

(2) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046–1053.

(3) *MACCS Structural keys*; Symyx Software: San Ramon, CA.

(4) James, C. A.; Weininger, D. *Daylight Theory Manual*; Daylight Chemical Information Systems Inc.: Irvine, CA, 2009.

(5) Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: Foundations, limitations and novel approaches. *Drug Discovery Today* **2007**, *12*, 225–233.

(6) Stumpfe, D.; Bajorath, J. Similarity Searching. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 260–282.

(7) Schneider, G.; Schneider, P.; Renner, S. Scaffold-Hopping: How Far Can You Jump?. *QSAR Comb. Sci.* **2006**, *25*, 1162–1171.

(8) Brown, J.; Jacoby, E. On Scaffolds and Hopping in Medicinal Chemistry. *Mini-Rev. Med. Chem.* **2006**, *6*, 1217–1229.

(9) Vogt, M.; Stumpfe, D.; Geppert, H.; Bajorath, J. Scaffold Hopping Using Two- Dimensional Fingerprints: True Potential, Black Magic, or a Hopeless Endeavor? Guidelines for Virtual Screening. *J. Med. Chem.* **2010**, *53*, 5707–5715.

(10) Gardiner, E. J.; Holliday, J. D.; O'Dowd, C.; Willett, P. Effectiveness of 2D Fingerprints for Scaffold Hopping. *Future Med. Chem.* **2011**, *3*, 405–414.

(11) Stumpfe, D.; Bill, A.; Novak, N.; Loch, G.; Blockus, H.; Geppert, H.; Becker, T.; Hoch, M.; Schmitz, A.; Kolanus, W.; Famulok, M.; Bajorath, J. Targeting Multi-Functional Proteins by Virtual Screening: Structurally Diverse Cytohesin Inhibitors with Differentiated Biological Functions. *ACS Chem. Biol.* **2010**, *5*, 839–849.

(12) Stumpfe, D.; Bajorath, J. Applied virtual screening: strategies, recommendations, and caveats. In *Methods and Principles in Medicinal Chemistry. Virtual Screening. Principles, Challenges, and Practical Guidelines*; Sotriffer, C., Ed.; Wiley-VCH: Weinheim, Germany, 2011; pp 73−103.

(13) Hert, J.; Willet, P.; Wilton, D. J. Comparison of Fingerprint-based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. J. *Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.

(14) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Enhancing the Effectiveness of Similarity-Based Virtual Screening Using Nearest-Neighbor Information. *J. Med. Chem.* **2005**, *48*, 7049–7054.

(15) Willett, P. Searching techniques for databases of two- and three-dimensional chemical structures. *J. Med. Chem.* **2005**, *48*, 4183–4199.

(16) Geppert, H.; Vogt, M.; Bajorath, J. Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205–216.

(17) Bender, A.; Glen, R. C. *MOLPRINT 2D*; Center for Molecular Science informatics, University of Cambridge: Cambridge, U.K.; http://www.molprint.com/. Accessed October 1, 2009.

(18) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naïve Bayesian Classifier. *J. Chem. Inf. Model.* **2004**, *44*, 170–178.

(19) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(20) Nisius, B.; Vogt, M.; Bajorath, J. Development of a Fingerprint Reduction Approach for Bayesian Similarity Searching Based on Kullback−Leibler Divergence Analysis. *J. Chem. Inf. Model.* **2009**, *49*, 1347–1358.

(21) Nisius, B.; Bajorath, J. Fingerprint Recombination − Generating Hybrid Fingerprints for Similarity Searching from Different Fingerprint Types. *ChemMedChem* **2009**, *4*, 1859–1863.

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

(23) Cover, T. M.; Thomas, J. A. *Elements of Information Theory*; Wiley: New York, 1991.

(24) Quinlan, J. R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann: San Mateo, CA, 1993.

(25) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities. *Nucleic Acids Res.* **2007**, *35*, D198–D201.

(26) *Scitegic Pipeline Pilot*; Accelrys, Inc.: San Diego, CA, 2010.

(27) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(28) Xu, Y.-J.; Johnson, M. Using Molecular Equivalence Numbers to Visually Explore Structural Features that Distinguish Chemical Libraries. *J. Med. Chem.* **2002**, *42*, 912–926.

(29) 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.