Journal of
Medicinal
Chemistry
Article

DOI: 10.1021/jm100492z

Scaffold Hopping Using Two-Dimensional Fingerprints: True Potential, Black Magic, or a Hopeless Endeavor? Guidelines for Virtual Screening

Martin Vogt, Dagmar Stumpfe, Hanna Geppert, and Jürgen Bajorath*

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

Received April 21, 2010

The scaffold hopping potential of popular 2D fingerprints has been thoroughly investigated. We have found that these types of fingerprints have at least limited scaffold hopping ability including early enrichment of small numbers of active scaffolds at high database ranks. However, it has not been possible to derive Tanimoto coefficient value ranges for individual fingerprints that are generally preferred for scaffold hopping. For selected fingerprints, similarity threshold values have been identified that yield small database selection sets having a high probability to contain a few active scaffolds. Furthermore, essentially all tested fingerprints have shown the ability to enrich scaffold hops in approximately 1% of a screening database. For the test cases reported herein, selecting 0.5−1% of the screening database yields ∼25% of the available scaffolds. On the basis of our findings, practical guidelines for virtual screening using different types of 2D fingerprints have been formulated.

Introduction

For ligand-based virtual screening (LBVS[a]), a variety of computational approaches have been introduced that can roughly be divided into similarity search and compound classification methods.[1] Popular LBVS methods and molecular representations that are utilized often greatly vary in their level of sophistication. Among relatively simplistic search tools are fingerprints calculated from two-dimensional (2D) molecular graph representations, so-called 2D fingerprints.[2,3] Typically, these fingerprints are bit string representations of molecular structure and properties. Various 2D fingerprint designs have been introduced that encode different types of molecular descriptors.[2,3] Despite their simplicity, 2D fingerprints are widely used for chemical similarity searching and virtual screening for bioactive compounds, although they generally do not encode structure−activity relationship (SAR) information. In both benchmark trials and practical applications, 2D fingerprints have often been found to be surprisingly successful in retrieving active compounds.[4−8]

Because fingerprints do not directly capture SAR information, the relationships between chemical similarity and activity similarity must be established, which is not a trivial task.[3] Fingerprint similarity searching generates a database ranking by quantifying fingerprint overlap using similarity coefficients and utilizing the resulting values as a measure of molecular similarity.[9] To what extent so derived molecular similarity correlates with activity similarity or, in other words, which similarity values are a reliable indicator of similar biological activity is currently unknown for most fingerprints. In a seminal study analyzing neighborhood behavior, scientists from Tripos demonstrated, among other aspects, that for their UNITY fingerprints, a Tanimoto coefficient (Tc)[9] value of at least 0.85 indicated a high probability that the compared compounds had the same bioactivity.[10] Although neighborhood behavior is fingerprint- and similarity coefficient-dependent,[3] this Tc value of 0.85 has been propagated in the literature as a general threshold for bioactivity, although it has been shown that it is not reliable when other fingerprints are utilized.[11] Hence, much work will be required to establish similarity threshold values of bioactivity for other fingerprints, if they in fact exist. From this point of view, fingerprint searching for active compounds is still in its infancy despite its long history and its popularity.

Of course, one might argue, rightfully so, that the most similar compounds detected in a similarity search should always have the highest probability to be active. Unfortunately, these are potentially active compounds one is usually not very interested in because they tend to be close structural analogues of known active reference compounds. Rather, "scaffold hopping" is the primary goal.[12−18] Scaffold hopping refers to the search for active compounds having different core structures than known reference compounds. Thus, one deliberately attempts to depart from known active chemotypes and identify structurally diverse active compounds, rather than analogues. Again, in this context, active compounds are typically considered diverse if they have unique core structures.

In particular for 2D fingerprints, scaffold hopping potential has remained controversial.[13] Proponents of more sophisticated LBVS approaches might occasionally question the principal ability of rather simple fingerprints to recognize remote similarity relationships. However, simplicity is probably not the main reason for questioning the scaffold hopping

*To whom correspondence should be addressed. Phone: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

[a] Abbreviations: LBVS, ligand-based virtual screening; SAR, structure−activity relationship; Tc, Tanimoto coefficient; 2D, two-dimensional; MACCS, Molecular ACCess System structural keys; ECFP, Extended Connectivity FingerPrint; ECFC, Extended Connectivity Fingerprint with Counts; TGT, Typed-Graph Triangles; GpiDAPH3, Graph-Π-Donor−Acceptor-Polar-Hydrophobe-Triangle.

© 2010 American Chemical Society

Published on Web 07/15/2010

Table 1. Scaffold Data Sets

target ID	target name	no. molecules	no. scaffolds
EB1-9	epidermal growth factor receptor erbB1	55	11
CA2-15	carbonic anhydrase II	70	14
D2R-72	dopamine D2 receptor	95	19
CB1-87	cannabinoid CB1 receptor	80	16
GRH-118	gonadotropin-releasing hormone receptor	65	13
SRT-121	serotonin transporter	80	16
KOR-137	κ opioid receptor	65	13
ECB-174	estrogen receptor β	70	14
HIP-191	human immunodeficiency virus type 1 protease	165	33
CFX-194	coagulation factor X	85	17
HIR-228	human immunodeficiency virus type 1 reverse transcriptase	100	20
N1R-250	neurokinin 1 receptor	60	12
AA3-280	adenosine A3 receptor	130	26
MR4-10142	melanocortin receptor 4	100	20
DHF-10457	dihydrofolate reductase	65	13
VEG-10980	vascular endothelial growth factor receptor 2	160	32
MCH-19905	melanin-concentrating hormone receptor 1	230	46

potential of 2D fingerprints (or fingerprints in general). Rather, similarity-based ranking becomes again a focal point. If structurally diverse active compounds are detected in a fingerprint search, they will typically not appear at top ranks but at lower rank positions because ranking is based on structural similarity, not activity similarity. Where to look for structurally diverse active compounds in database rankings is another unsolved problem in fingerprint similarity searching, in addition to the lack of generally applicable activity similarity, threshold values, as discussed above. In fact, to our knowledge, Tc value ranges where scaffold transitions among active compounds preferentially occur have not yet been determined for any standard fingerprint, although such insights would be of prime importance for practical similarity search application. It is not even clear if preferred similarity value ranges for scaffold hopping exist at all and, if so, how they might compare across different fingerprint designs and compound classes. Hence, in practice, in order to increase the probability of successfully selecting structurally diverse active compounds from database rankings, information from multiple search trials using different fingerprint methods is often combined.[19] Alternatively, investigators must have the ability or intuition to cherry-pick active compounds from rankings, hence rendering compound selection a form of art, if not "black magic".

Moreover, assessing scaffold hopping potential, i.e., the ability to retrieve structurally diverse active compounds in similarity searching, is further complicated by the fact that no generally accepted definition currently exists of what constitutes a successful scaffold transition.[20] In many benchmark investigations, the term scaffold is rather loosely used and it often remains unclear how scaffold hopping potential has exactly been evaluated.[20]

In light of this situation, given the importance of scaffold hopping guidelines for fingerprint similarity searching, we decided to systematically analyze the scaffold hopping potential of conventional 2D fingerprints and characterize their search behavior. To these ends, we have generated a carefully designed scaffold hopping benchmark system as the basis for our analysis. This compound database is made publicly available to support further investigations. Five conventional fingerprints representing different designs have been systematically analyzed and compared in calculations applying established similarity search strategies. We show that it is generally difficult, but not impossible, to find structurally diverse active compounds in small database selection sets of fewer than 100 compounds. However, we also show that 2D fingerprints generally enrich different bioactive scaffolds in approximately 1% of the screening database, which has clear implications for how fingerprint similarity searching is best applied. Taken together, the findings presented herein have made it possible to formulate some general guidelines how to best utilize different 2D fingerprints in virtual screening.

Materials and Methods

Scaffold Definition and Source. To consistently derive scaffolds for our analysis, we applied the scaffold definition of Bemis and Murcko.[21] Following this definition, scaffolds were extracted from compounds by removing all R-groups but retaining linkers between ring systems. This hierarchical fragmentation approach currently is the most widely applied and established scaffold definition. We did not consider unusual scaffolds, i.e., scaffolds that contained rings with more than six atoms or poly amine structures. As a scaffold source, we utilized the freely available ChEMBL database[22] that contains curated sets of active compounds mostly originating from pharmaceutical sources.

Bemis and Murcko scaffolds were also transformed into "carbon skeletons" and "reduced cyclic skeletons" according to Xu and Johnson.[23] Carbon skeletons are derived from scaffolds by changing each heteroatom to a carbon atom and all bond orders to single bonds. Thus, different carbon skeletons represent topologically distinct scaffolds. Reduced cyclic skeletons further abstract from carbon skeletons by ignoring differences in ring size and linker length (i.e., all rings are of the same size and linkers have unit length).

Data Set Selection. For scaffold derivation, only compounds were considered with potency (K_i or IC_{50} value) of at least 1 μM. Molecular size was restricted to a maximum of 50 non-hydrogen atoms. Furthermore, to ensure that there was no unreasonably large discrepancy between compound and scaffold size, compounds representing a given scaffold were only selected if their R-groups had in total not more non-hydrogen atoms than the scaffold. A target-directed scaffold set had to contain a minimum of 10 distinct scaffolds, each of which had to be represented by five different compounds/analogues. On the basis of these criteria, we derived compound and scaffold sets for 17 different compound activity classes containing a total of 1675 unique compounds comprising 334 unique scaffolds (335 scaffolds in total), as reported in Table 1. The activity classes contained between 55 and 230 compounds corresponding to 11−46 scaffolds, i.e., each scaffold was represented by exactly five compounds. For compounds and scaffolds, the number of heavy atoms ranged from 11 to 49 and 6 to 37, respectively, and the molecular weight from 166.2 to 708.6 Da and 78.1 to 515.8 Da, respectively. Upon publication, these compound/scaffold data sets are freely available via the following URL: http://www.lifescienceinformatics.uni-bonn.de.

Screening Databases. As background databases for similarity search calculations, ChEMBL and a subset of ZINC[24] were selected. After removal of all compounds with activity annotations for our 17 targets, a total of 492415 ChEMBL compounds remained (all of which with known activity) that were used as one of two screening databases. Furthermore, from ZINC, a total of 507594 compounds (with largely unknown activity annotations) were selected that fell into the same molecular weight and heavy atom ranges defined by our test compounds. Two background databases were used for fingerprint similarity searching as a control for database-dependent effects on search results.

Fingerprints. We selected five representative 2D fingerprints of different design that are available in the popular Pipeline Pilot[25] and/or Molecular Operating Environment[26] chemoinformatics platforms including Molecular ACCess System (MACCS) structural keys[27] (166 bit positions), Extended Connectivity FingerPrint with bond diameter 4 (ECFP4; ~4 × 10^9 possible features),[25] Extended Connectivity Fingerprint with Counts and bond diameter 4 (ECFC4; ~4 × 10^9 possible features),[25] the Typed-Graph Triangles (TGT; 1704 bits),[26] and the Graph-Π-Donor−Acceptor-Polar-Hydrophobe-Triangle (GpiDAPH3; 30240 possible features)[26] fingerprint. MACCS consists of 166 structural substructures/patterns with 1−10 non-hydrogen atoms. ECFP4 is a combinatorial molecule-specific fingerprint that encodes layered atom environments with a maximum diameter of four bonds around each atom in a molecule, and ECFC4 is the corresponding nonbinary count fingerprint. This means that it not only detects unique connectivity patterns in a molecule, like ECFP4, but also records how often each feature is generated. TGT and GpiDAPH3 fingerprints are pharmacophore fingerprints calculated from 2D molecular graphs that account for three-point pharmacophore patterns. TGT assigns each atom to one of four atom types (hydrogen bond donor or base, hydrogen bond acceptor or acid, both hydrogen bond acceptor and donor, or hydrophobic) and interatomic distances are divided into six different bond distance ranges. For GpiDAPH3, each atom is assigned to one of eight types derived from three atomic properties ("in π system", "is donor", and "is acceptor") and interatomic distances are divided into eight different bond distance ranges.

These five fingerprints were selected because they represent structural fragment, pharmacophore, or topological fingerprints, which are major 2D fingerprint categories. Furthermore, they are of distinct design and have different format, complexity, and size. For similarity calculations using the nonbinary ECFC4 fingerprints, the general form of the Tanimoto coefficient was used[9] and for the other four fingerprints, binary Tc calculations were carried out.

Similarity Searching. For fingerprint searching, all five compounds representing an individual scaffold were used once as reference molecules to search for compounds representing the remaining scaffolds (10−45) for each activity class. Hence, for each activity class, between 11 and 46 search trials were carried out (see Table 1). For these calculations, all active test compounds (except the respective reference set) were added to the background databases. For similarity searching, the one nearest neighbor (1-NN) search strategy[4] was applied. In 1-NN calculations, the final similarity value for a database compound relative to the reference set is obtained by selecting the maximum observed Tc value between the database compound and each reference molecule. This search strategy often favors the detection of structurally similar compounds (for example, analogues of reference compounds)[5] and hence is expected to yield a lower limit for (and thus conservative assessment of) the scaffold hopping potential of fingerprints in similarity search calculations using multiple reference compounds.

The results of fingerprint search calculations were monitored for each activity class and each scaffold and averages were also calculated. "Scaffold recall" was determined as the number of distinct scaffolds retrieved within the top 100, 500, 1,000, and 5000 database compounds (corresponding to approximately 0.02%, 0.1%, 0.2%, and 1.0% of the screening database, respectively). Furthermore, for each fingerprint, the distribution of Tc values was determined for database and active test compounds and, in addition, rank positions were analyzed. For database compounds, the median rank for a given Tc value was calculated. Recall rates were also calculated for carbon skeletons and reduced cyclic skeletons.

Results and Discussion

Our aims have been to systematically assess the scaffold hopping potential of contemporary 2D fingerprints, investigate Tc value ranges where scaffold hops might occur, and compare similarity value distributions and ranks among active and database compounds. To these ends, we have designed a compound benchmark system that has made it possible to evaluate scaffold hopping potential in a well-defined manner. We have also ensured that active and screening database compounds had comparable size in order to avoid similarity search bias through molecular size effects.[28] In the systematic search calculations reported herein, any detection of a true-positive active compound represented a successful scaffold transition from one Bemis and Murcko scaffold to another because we exclusively used sets of reference compounds whose common scaffold was not contained in any other active compound. Furthermore, for any of our activity classes, at least 11 distinct scaffolds were available as targets for similarity searching, each of which was represented by five different compounds. Initially, we determined global scaffold (and skeleton) recall rates for different 2D fingerprints.

Global Scaffold Recall. In Table 2, we report the results of systematic similarity search calculations over all activity classes for the ChEMBL and ZINC background databases and selection sets of 100 and 5000 compounds. For the top 100 database compounds, scaffold recall rates for the different fingerprints ranged from 8.4 to 13.5% for ChEMBL and 13.9 to 21.9% for the ZINC background database. For the top 5000 ranks, i.e., 1% of the screening database, recall rates ranged from 28.2 to 41.7% (ChEMBL) and 29.6 to 42.8% (ZINC). Five general observations were made. First, each of the five fingerprints showed at least limited scaffold hopping ability. Second, on average, scaffold recall in small selection sets did not substantially differ for the studied fingerprints, although they were of different design and in part very different complexity. Third, as indicated by relatively large standard deviations (Table 2), significant fluctuations between individual search trials were observed, as discussed in more detail below. These large standard deviations are a general characteristic of fingerprint search calculations reflecting the influence of selecting different reference compound sets, especially because the reference sets used herein are composed of compounds sharing the same scaffold (i.e., different reference sets represent distinct scaffolds). This represents another complication of similarity searching. Fourth, scaffold recall rates were overall slightly higher for the ZINC than the ChEMBL database, probably because ChEMBL compounds are on average more similar to each other than ChEMBL compared to ZINC compounds (and hence more difficult to differentiate). However, the relative fingerprint performances for these screening databases were equivalent. Because corresponding trends were observed, we

Table 2. Recall Rates for Scaffolds, Carbon Skeletons, And Reduced Cyclic Skeletons[a]

	ChEMBL					
	100			5000		
	B-M	carbon Sk	reduced Sk	B-M	carbon Sk	reduced Sk
MACCS	10.2 (7.2)	6.2 (5.5)	3.6 (4.6)	31.5 (15.1)	25.7 (13.2)	22.8 (15.2)
ECFC4	13.5 (8.0)	9.9 (7.5)	7.0 (7.7)	36.1 (17.2)	30.9 (16.5)	26.9 (21.5)
ECFP4	13.5 (7.3)	10.5 (6.6)	6.9 (6.1)	41.7 (21.9)	36.8 (21.3)	34.0 (25.1)
GpiDAPH3	13.3 (8.0)	10.0 (7.0)	6.5 (6.7)	37.0 (18.7)	32.6 (19.6)	29.2 (24.5)
TGT	8.4 (6.2)	5.7 (4.8)	2.9 (3.4)	28.2 (18.0)	22.8 (17.9)	20.1 (20.5)
	ZINC					
	100			5000		
	B-M	carbon Sk	reduced Sk	B-M	carbon Sk	reduced Sk
MACCS	14.5 (11.0)	10.0 (8.5)	7.3 (7.5)	35.9 (15.6)	30.3 (14.2)	27.1 (15.1)
ECFC4	20.9 (14.2)	16.5 (14.6)	13.8 (17.3)	38.7 (17.8)	33.7 (17.9)	30.1 (22.4)
ECFP4	21.6 (13.5)	17.4 (12.4)	13.2 (14.8)	42.8 (22.6)	38.7 (22.9)	37.1 (27.8)
GpiDAPH3	21.9 (14.4)	17.9 (12.8)	15.2 (15.7)	38.8 (19.7)	34.3 (21.2)	30.2 (24.5)
TGT	13.9 (14.1)	10.8 (12.3)	7.5 (9.6)	29.6 (19.3)	24.8 (19.4)	21.4 (23.3)

[a] Average recall rates (in %) are reported for Bemis and Murcko scaffolds (B-M), carbon skeletons (carbon Sk), and reduced cyclic skeletons (reduced Sk) over all activity classes for the top-ranked 100 and 5000 database compounds. Standard deviations for individual search trials are given in parentheses.

discuss in the following representative and average search results obtained for the ChEMBL background database, shown in Figures 1−4, and provide complete results for ChEMBL and the corresponding results for the ZINC background database in Supporting Information Figures S1−S3 and Supporting Information Table S1. Fifth, Table 2 also shows that recall rates comparable in magnitude to scaffolds were also observed for carbon skeletons and reduced cyclic skeletons that represent further abstractions from scaffolds at varying levels. Because multiple scaffolds might correspond to the same carbon skeleton and multiple carbon skeletons to the same reduced cyclic skeleton, recall rates typically decrease in this order. However, the observation that their magnitude was similar for essentially all fingerprints demonstrates that the majority of detected scaffolds were topologically distinct. Hence, in the following, we focus on the analysis of scaffold recall characteristics.

Data Representation. To comprehensively analyze relationships between scaffold recall/hopping, corresponding Tc value ranges, and database rank positions, we designed an information-rich data representation for comparison of individual search trials, as shown in Figure 1. At first glance, this representation is fairly complex and, therefore, Figure 1a provides a representative example with detailed explanations of the elements of data display and analysis. For individual search trials and database ranks of up to 500000 molecules, correctly identified active compounds representing a scaffold hop are shown with their rank positions and corresponding color-coded Tc values. Furthermore, average cumulative ranks for recall of 25%, 50%, and 75% of all scaffolds in an activity class are reported. This data representation provided the basis for our subsequent analysis.

Fingerprint Search Phenotypes. The representative search trials in Figure 1 illustrate in part rather general search and scaffold hopping characteristics of the compared fingerprints. For example, Figure 1a reveals that MACCS search calculations enrich small numbers of active scaffolds at high rank positions and high Tc values of > 0.8 (red shaded area), a trend that has been observed for many different activity classes. As reported in Table 2, the overall scaffold recall for the top 100 database ranks achieved with MACCS was 10.2%, which is quite consistent with the results shown in Figure 1a. Typically, however, the majority of scaffold hops were found at lower Tc values of 0.4−0.6 (green shaded area) together with more than 100000 database compounds. Hence, in these cases, most active scaffolds vanish in the background database. For the example in Figure 1a, in order to achieve 25% scaffold recall with MACCS, the first approximately 1000 database ranks had to be selected. A scaffold recall rate of at least 25% was generally considered a lower limit for successful scaffold hopping and retrieval.

For the individual trials shown in Figure 1a, scaffold hops were underrepresented within the Tc interval 0.6−0.8, but these observations varied from class to class. For the different reference sets, there was only little variation in observed Tc distributions, i.e., similar Tc values corresponded to similar database ranks, in contrast to ECFC4 in Figure 1b, but comparable to ECFP4 in Figure 1c. ECFC4 is the count fingerprint variant of ECFP4, and its individual search trials were characterized by highly variable Tc profiles. However, despite this variability, a fairly constant distribution of scaffold hops over almost the entire Tc range was observed including high ranks (with, on average, 13.5% scaffold recall for the top 100 database compounds; Table 2). ECFP4 produced overall much lower Tc values than ECFC4. As can be seen in Figure 1c, there was a clear enrichment of scaffold hops within the Tc interval 0.2−0.4 (blue shaded area), active scaffolds were similarly distributed over this interval, and there was only little variation over individual search trials, similar to MACCS. Such observations were also made for other activity classes. In the example shown for ECFP4, only approximately 100 database ranks needed to be considered in order to achieve 25% scaffold recall. The other fingerprints performed similarly well on this specific activity class. Different from ECFP4, the pharmacophore-type fingerprints GpiDAPH3 (Figure 1d) and TGT (Figure 1e) displayed highly variable Tc distributions and significantly different scaffold recall in individual trials using different reference sets. In both cases, many scaffolds congregated in very low Tc ranges, i.e., 0.0−0.2 (purple shaded area) for GpiDAPH3 and 0.2−0.4 (blue shaded area) for TGT, together with 10000−100000 or more (for TGT) database compounds. For GpiDAPH3, the Tc range 0.0−0.2

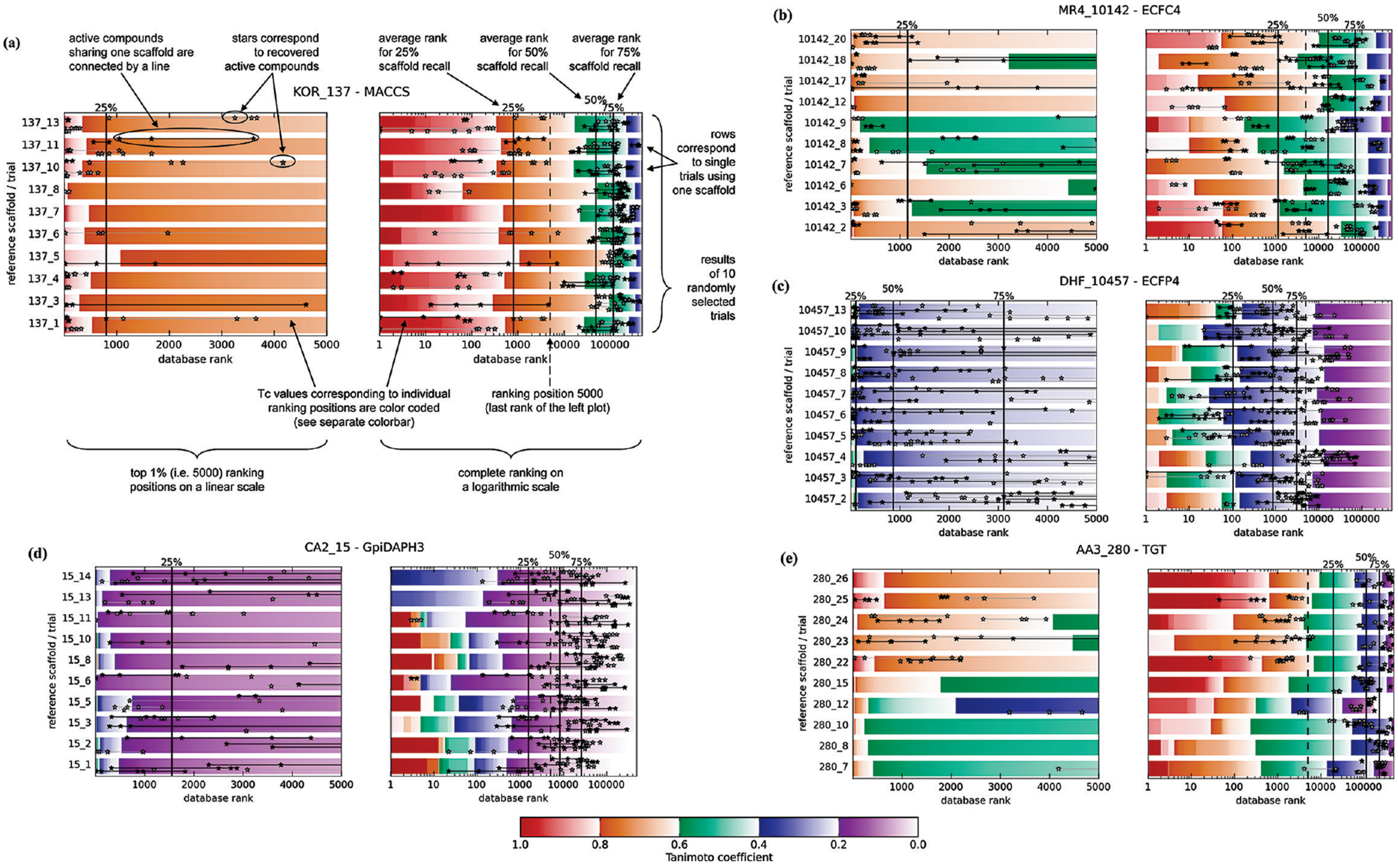

Figure 1. Representative examples of individual search trials. For exemplary individual search trials, relationships between the detection of active compounds, Tc values, and ranking positions are visualized for the different fingerprints: (a) MACCS, (b) ECFC4, (c) ECFP4, (d) GpiDAPH3, (e) TGT. To aid in the interpretation of these information-rich representations, (a) contains detailed descriptions. For each fingerprint, 10 randomly selected search trials for one activity class are shown that correspond to the individual rows in the panels. The left panel provides information for the top 1% of the database ranking (i.e., 5000 ranks) using a linear scale for ranking positions. By contrast, the panel on the right shows complete database rankings on a logarithmic scale. Ranking positions of active compounds are marked using black or gray stars. Active compounds containing the same scaffold are connected. Tc values corresponding to individual ranking positions are reported using a color code (see the color bar at the bottom of the figure): 1.0−0.8, red; 0.8−0.6, orange; 0.6−0.4, green; 0.4−0.2, blue; 0.2−0.0, violet. Solid vertical lines mark ranking positions that correspond to 25%, 50%, or 75% scaffold recall when averaging the search results over all trials for one activity class and fingerprint. The dashed vertical line in the right panel marks the database ranking position 5000. In this and the following figures, results are reported for the ChEMBL background database.

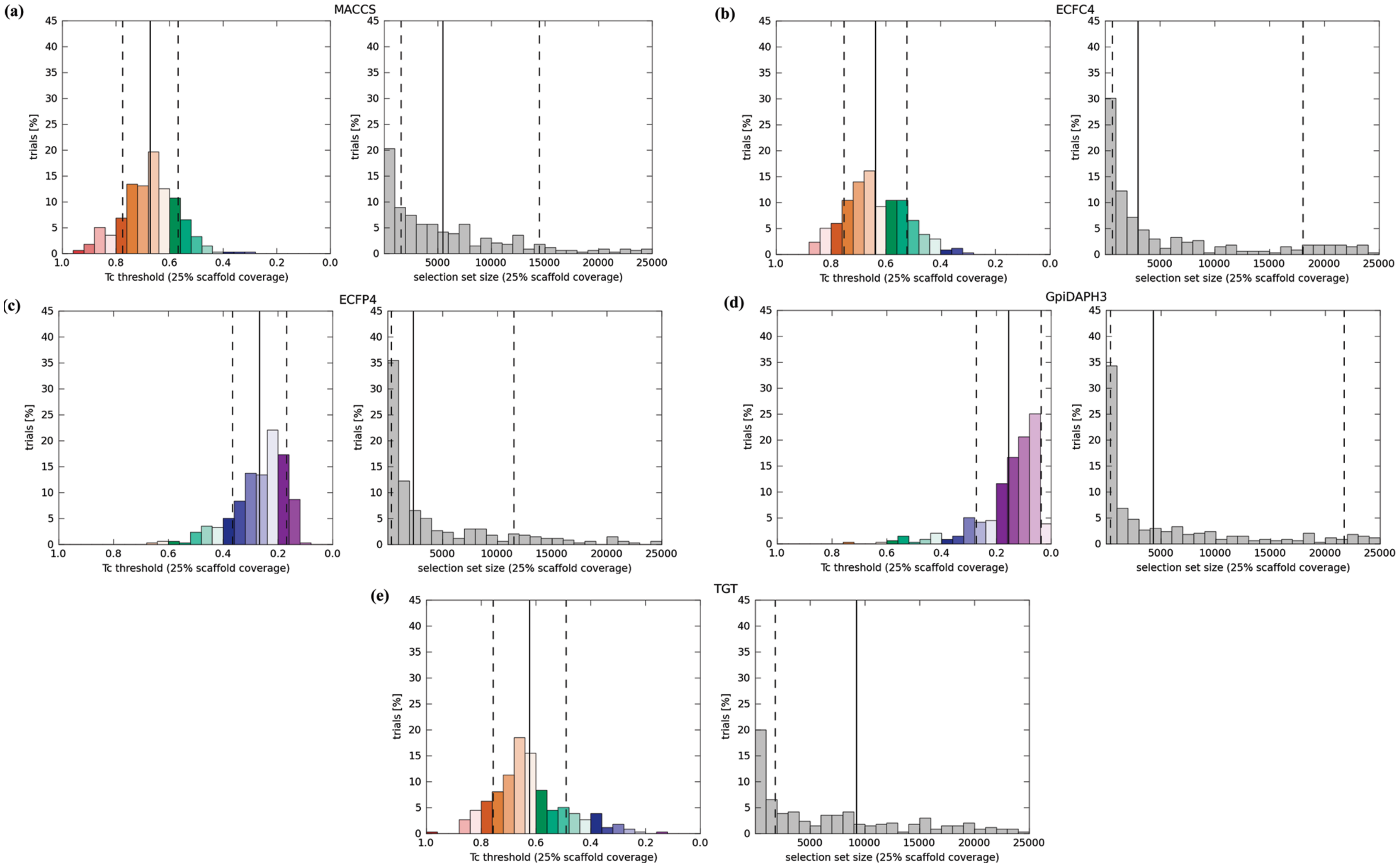

Figure 2. Scaffold recall: Tc threshold values and database selection sets. Reported are the distribution of Tc threshold values (left) and corresponding database selection set sizes (right) that recover 25% of the scaffolds of a compound class. For each individual fingerprint, the results are summarized for all activity classes and search trials. On the left, the number of trials is reported for each Tc threshold value at which at least one compound for 25% of all scaffolds is recovered. These Tc values vary for individual trials. The solid line shows the mean of the Tc threshold distribution and the dashed lines show the standard deviation. On the right, the distribution of the database selection set sizes required to achieve a scaffold recall rate of 25% is reported (up to a selection set size of 25000 database compounds). In this case, the solid line shows the median of the distribution while the dashed lines mark the interquartile range. (a) MACCS, (b) ECFC4, (c) ECFP4, (d) GpiDAPH3, (e) TGT.

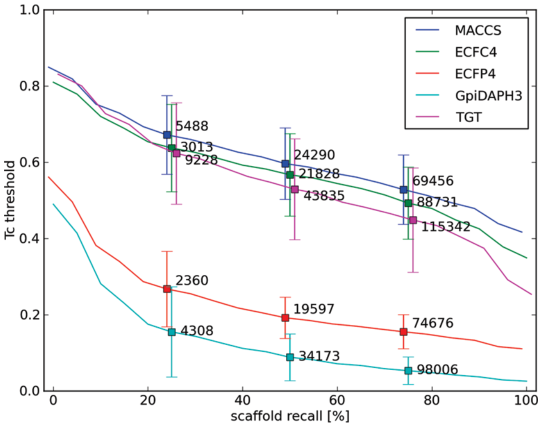

Figure 3. Average Tc threshold values for scaffold recall rates. For MACCS (blue), ECFC4 (green), ECFP4 (red), GpiDAPH3 (cyan), and TGT (magenta), the average Tc threshold value required to achieve a given scaffold recall rate is reported. The variation in these Tc values across all trials (error bars) is shown for recall rates of 25%, 50%, and 75%. Numbers next to the error bars report the median database selection set size for which a recall rate is achieved.

consistently was the most populated interval across different activity classes. On the other hand, TGT showed much greater variation in Tc ranges across classes and reference sets, covering a Tc range from about 0.2−0.7. Despite the variability of individual search trials, GpiDAPH3 produced a constant limited early enrichment of scaffold hops. On average, TGT displayed the weakest search performance. Importantly, Figure 1 also reveals that different fingerprints enriched active scaffolds at very different Tc values, e.g., MACCS > 0.8, ECFP4 > 0.2, or GpiDAPH3 predominantly < 0.2. Consequently, it is essentially impossible to define generally applicable Tc threshold values for scaffold hopping.

Similarity Value Distributions and Database Selection Sets. Next we investigated the general distribution of Tc threshold values over all activity classes for which a scaffold recall of at least 25% was achieved and determined the corresponding database selection set sizes. The results are shown in Figure 2. For all fingerprints, Tc value ranges that detected at least 25% of the possible scaffold hops for diverse active compounds were found to vary substantially. The Tc value range distributions of MACCS (Figure 2a) and ECFC4 (Figure 2b) were comparable, but the distribution of ECFC4 was

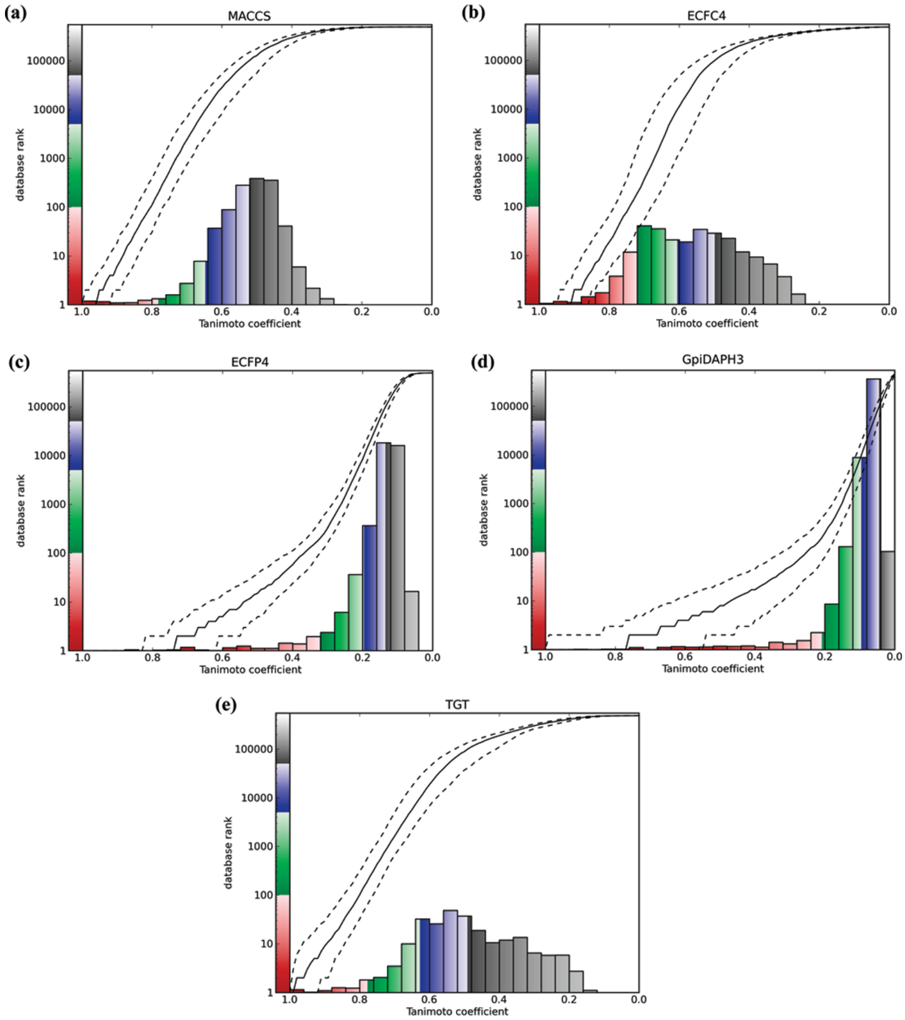

Figure 4. Similarity values for scaffold hops and corresponding median ranks. Distributions of Tc values of active compounds (scaffold hops) across all trials are shown. The bars are color-coded according to the median database rank corresponding to the Tc values. The red color range corresponds to ranks 1−100, the green to ranks 100−5000, and the blue to range from 5000 to 50000. The solid curve monitors Tc values and the corresponding median ranks and the dashed lines show the interquartile range, i.e. the variation in ranks observed over all trials. (a) MACCS, (b) ECFC4, (c) ECFP4, (d) GpiDAPH3, (e) TGT.

somewhat narrower. By contrast, the distributions of ECFP4 (Figure 2c) and, in particular, GpiDAPH3 (Figure 2d), were considerably shifted toward small Tc values, whereas the distribution of TGT (Figure 2e) displayed overall broadest coverage of the entire Tc range. The Tc value ranges where a notable number of scaffold hops for different compound classes occurred were highly variable. Thus, for these fingerprints, it was not possible to determine narrow and generally applicable Tc values where scaffold hops preferentially occurred. Database selection set sizes corresponding to 25% scaffold recall also varied for different trials, as revealed in Figure 2.

Scaffold Recall Potential. Calculating median values of these distributions made it possible to compare the scaffold recall potential of different fingerprints. In decreasing order of search performance, for ECFP4, ECFC4, GpiDAPH3, MACCS, and TGT, median database selection set sizes of approximately 2400, 3000, 4300, 5500, and 9200 compounds, respectively, were required to obtain 25% of all active scaffolds. Hence, with the exception of TGT, a significant enrichment of scaffolds was achieved by selecting approximately 0.5−1% of the screening database. This analysis is further extended in Figure 3 that reports average Tc threshold values for increasing scaffold recall and median database selection set sizes for 25%, 50%, and 75% recall. To achieve 50% scaffold recall, the top three fingerprints at this level, ECFP4, ECFC4, and MACCS, required selection of approximately 19000−24000 database compounds. For 75% recall, at least approximately 69000 database compounds needed to be selected (MACCS). Hence, while a notable general enrichment of active scaffolds was observed in 0.5−1% top-ranked database compounds for four of five fingerprints, complete (or nearly complete) scaffold coverage could not be achieved through fingerprint searching. In addition, Figure 3 also highlights much lower scaffold hopping-relevant Tc values observed for ECFP4 and GpiDAPH3 compared to the other fingerprints.

Database Ranks and Similarity Threshold Values. In Figure 4, the distribution of Tc values for active compounds is compared to corresponding database ranks. Although it was not possible to identify generally preferred Tc value ranges for scaffold hopping, as discussed above, in some instances, we were able to determine Tc threshold values for the enrichment of scaffolds at high database ranks. Essentially all five fingerprints displayed an early enrichment of at least a few active scaffolds. However, the definition of Tc threshold values for highly ranked scaffold hops was only meaningful in cases where the underlying distribution of Tc values was relatively narrow and reference set- and target-dependent variations in search performance were limited. For example, this has been the case for MACCS (Figure 4a) and ECFP4 (Figure 4c), but not ECFC4 (Figure 4b) or TGT (Figure 4e). For GpiDAPH3, the Tc distribution was narrow, but most of the actives produced Tc values < 0.2, whereas individual scaffold hops were observed at Tc values of up to 0.9, thus prohibiting the definition of meaningful threshold values. However, this was accomplished for MACCS and ECFP4 that detected a limited number of scaffold hops within the top 10−100 database compounds at Tc values above 0.8 and 0.4, respectively.

Guidelines for Virtual Screening. Taken together, the results of our analysis have made it possible to extrapolate to virtual screening situations and formulate some "rule-of-thumb" guidelines for the use of these fingerprints in practical applications.

MACCS. The bulk of scaffold hops occur within the Tc range 0.4−0.6 that usually also covers more than 50% of the screening database, which is not suitable for practical applications. Compounds with Tc values above 0.8 can be evaluated for a limited number of scaffold hops. A threshold value of 0.8 consistently yields database selection sets of approximately 100 compounds (for source database of ~500000 molecules).

ECFP4. This fingerprint showed overall the highest scaffold hopping potential in our study. A Tc threshold value of 0.4 can be applied to evaluate potential scaffold hops in database selection sets of approximately 100 compounds. Furthermore, a Tc threshold value of 0.2 can be applied to select approximately 1% of the database that is likely to contain a significant number of scaffold hops. These scaffold hops are widely distributed over the top 1% of the database ranks, suggesting random sampling of candidate compounds in this range or experimental screening of this small subset.

ECFC4. The scaffold recall performance is generally similar to ECFP4, but much higher variation in Tc value ranges are observed, hence prohibiting the definition of reliable similarity threshold values. Accordingly, compound selection should generally be based on database rank positions, rather than Tc values.

TGT. This fingerprint shows the lowest scaffold hopping potential in our study. Application in virtual screening would generally be problematic due to large variation in Tc value ranges and highly variable performance, depending on both reference sets and compound classes.

GpiDAPH3. This fingerprint displays an extreme Tc value distribution; nearly all database compounds consistently yield values < 0.2, which is not suited for derivation of Tc threshold values. The compound class-dependent variation in search performance is high, but the fingerprint is likely to substantially enrich diverse scaffolds within the top 1% of the database.

Conclusions

Representative fingerprints derived from 2D molecular representations of different design and complexity have been thoroughly investigated for their ability to detect compounds having similar activity but distinct scaffolds. For this purpose, a well-structured scaffold hopping benchmark system has been generated. On the basis of systematic search calculations, we have found that the fingerprints we analyzed had at least limited scaffold hopping potential although they differed in their search behavior. It has not been possible to determine generally preferred similarity value ranges for scaffold hopping, due to large variations in Tc values obtained for different reference sets and/or compound classes. Accordingly, compound selection on the basis of rank positions was generally preferred to Tc-based selection. Nevertheless, for at least two fingerprints, ECFP4 and MACCS, it has been possible to define Tc threshold values that can be applied to select small database selection sets of approximately 100 compounds displaying an early enrichment of scaffold hops (although only small numbers of alternative scaffolds might be available). However, essentially all five fingerprints are capable of enriching database subsets with active scaffolds; in four cases, selecting the top-ranked 0.5−1% of the screening database has been sufficient to retrieve on average 25% of the scaffolds belonging to 17 different activity classes. These findings have implications for practical similarity search applications using these types of 2D fingerprints. If the primary goal is the identification of only a few novel hits, the results of our analysis suggest that it is well worth analyzing the approximately top 100 database compounds. However, this might not be the preferred strategy for similarity searching. On the basis of our findings, it would often make more sense to preselect approximately 1% of the database for a subsequent limited screening campaign, thereby exploiting a likely notable enrichment of structurally diverse hits in a small subset of the database. In light of the results presented herein, we suggest that preselection of database subsets of such size (i.e., ~1%) presents a more meaningful application of fingerprint similarity searching than the "needles in haystacks" scenario associating with focusing on small numbers of top-ranked candidate molecules.

Supporting Information Available: Figures S1−S3 report complete individual and average fingerprint search results for the ChEMBL and ZINC background databases and correspond to Figures 1, 2, and 4 of the paper. Table S1 reports scaffold recovery rates for the top 100, 500, 1000, and 5000 compounds for ChEMBL and ZINC background databases corresponding to Table 2 of the paper. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Bajorath, J. Integration of virtual and high-throughput screening. *Nature Rev. Drug Discovery* **2002**, *1*, 882–894.

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

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

(4) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.

(5) Tovar, A.; Eckert, H.; Bajorath, J. Comparison of 2D fingerprint methods for multiple-template similarity searching on compound activity classes of increasing structural diversity. *ChemMedChem* **2007**, *2*, 208–217.

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

(7) Sheridan, R. P. Chemical similarity searches: when is complexity justified? *Expert Opin. Drug Discovery* **2007**, *2*, 423–430.

(8) 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, Vol. 48, in press.

(9) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

(10) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior−a useful concept for validation of molecular diversity descriptors. *J. Med. Chem.* **1996**, *39*, 3049–3059.

(11) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45*, 4350–4358.

(12) Brown, J.; Jacoby, E. On scaffolds and hopping in medicinal chemistry. *Mini. Rev. Med. Chem.* **2006**, *6*, 1217–1229.

(13) Stumpfe, D.; Bajorath, J. Similarity searching. In *Wiley Interdisciplinary Reviews: Computational Molecular Science*, in press.

(14) Renner, S.; Schneider, G. Scaffold-hopping potential of ligand-based similarity concepts. *ChemMedChem* **2006**, *1*, 181–185.

(15) Senger, S. Using Tversky similarity searches for core hopping: finding the needles in the haystack. *J. Chem. Inf. Model.* **2009**, *49*, 1514–1524.

(16) Tsunoyama, K.; Amini, A.; Sternberg, M. J. E.; Muggleton, S. H. Scaffold hopping in drug discovery using inductive logic programming. *J. Chem. Inf. Model* **2008**, *48*, 949–957.

(17) Wale, N.; Watson, I. A.; Karypis, G. Indirect similarity based methods for effective scaffold-hopping in chemical compounds. *J. Chem. Inf. Model* **2008**, *48*, 730–741.

(18) Barker, E. J.; Buttar, D.; Cosgrove, D. A.; Gardiner, E. J.; Gillet, V. J.; Kitts, P.; Willett, P. Scaffold-hopping using clique detection applied to reduced graphs. *J. Chem. Inf. Model.* **2006**, *46*, 503–511.

(19) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discovery Today* **2002**, *7*, 903–911.

(20) 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.

(21) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(22) ChEMBL, www.ebi.ac.uk/chembldb/index.php.

(23) 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.

(24) 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.

(25) *Scitegic Pipeline Pilot*; Accelrys, Inc.: San Diego, CA, 2009.

(26) *MOE* (*Molecular Operating Environment*); Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2007.

(27) *MACCS Structural Keys*; Symyx Software: San Ramon, CA, 2005.

(28) Wang, Y.; Bajorath, J. Advanced fingerprint methods for similarity searching: balancing molecular size effects. *Comb. Chem. High Throughput Screening* **2010**, *13*, 220–228.