JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

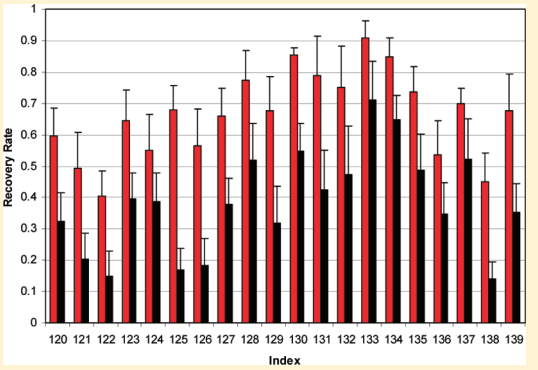
Large-Scale Similarity Search Profiling of ChEMBL Compound Data Sets

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: A large-scale similarity search investigation has been carried out on 266 well-defined compound activity classes extracted from the ChEMBL database. The analysis was performed using two widely applied two-dimensional (2D) fingerprints that mark opposite ends of the current performance spectrum of these types of fingerprints, i.e., MACCS structural keys and the extended connectivity fingerprint with bond diameter four (ECFP4). For each fingerprint, three nearest neighbor search strategies were applied. On the basis of these search calculations, a similarity search profile of the ChEMBL database was generated. Overall, the fingerprint search campaign was surprisingly successful. In 203 of 266 test cases (~76%), a compound recovery rate of at least 50% was observed with at least the better performing fingerprint and one search strategy. The similarity search profile also revealed several general trends. For example, fingerprint searching was often characterized by an early enrichment of active compounds in database selection sets. In addition, compound activity classes have been categorized according to different similarity search performance levels, which helps to put the results of benchmark calculations into perspective. Therefore, a compendium of activity classes falling into different search performance categories is provided. On the basis of our large-scale investigation, the performance range of state-of-the-art 2D fingerprinting has been delineated for compound data sets directed against a wide spectrum of pharmaceutical targets.



INTRODUCTION

Molecular fingerprints are usually defined as bit string representations of molecular structure and properties and have for more than two decades been utilized in chemical similarity searching and virtual screening for new active compounds.[1−3] Fingerprints can be classified into 2D and 3D molecular representations, dependent on molecular graph- or conformation-derived features that are utilized for their design.[1,4] Regardless of specific design criteria, fingerprint search calculations involve the comparisons of fingerprints calculated for reference and database compounds and the quantitative comparison of fingerprint (bit string) overlap as a measure of molecular similarity.[1] Accordingly, fingerprint searching is an intrinsically simple similarity method, especially when 2D fingerprints are used that only require the molecular graph as input. Various fingerprint engineering[4−6] and similarity search strategies[7−9] have been introduced to further improve fingerprint performance and/or tune fingerprints for compound class-specific search calculations. Despite their conceptual simplicity, 2D fingerprints have been shown to display significant scaffold hopping potential in benchmark trials[10,11] and practical applications.[12,13]

Virtual screening tools including 2D fingerprints are typically evaluated in retrospective benchmark investigations on activity classes taken from compound databases.[14−16] Popular source databases include, for example, the MDDR,[17] which is, however, license-restricted similar to other commercial database products and, hence, not generally available. Therefore, carefully designed data sets that are made publicly available[10,18] are highly relevant for method evaluation[10,19] as well as public domain compound repositories,[20−22] especially those that collect compound activity and optimization data from medicinal chemistry literature or patent sources.[21,22] These compound databases provide a sound basis for the generation of compound data sets that can be freely shared for method comparison.

Here we have carried out an unconventional fingerprint similarity search investigation on public domain compound data. Rather than comparing the search performance of 2D fingerprints on a limited number of selected activity classes, which is typically done,[10,15,16] we have extracted all compound data sets from ChEMBL[22] that were suitable for fingerprint test calculations in order to generate a similarity search profile of this database. ChEMBL currently is the largest publicly available repository of curated compound activity data taken from medicinal chemistry sources. Furthermore, rather than evaluating different fingerprints on ChEMBL data sets to compare details of their relative search performance, we selected two 2D fingerprints that represent the current performance spectrum of these search tools.

Received: May 5, 2011
Published: July 05, 2011

ACS Publications © 2011 American Chemical Society

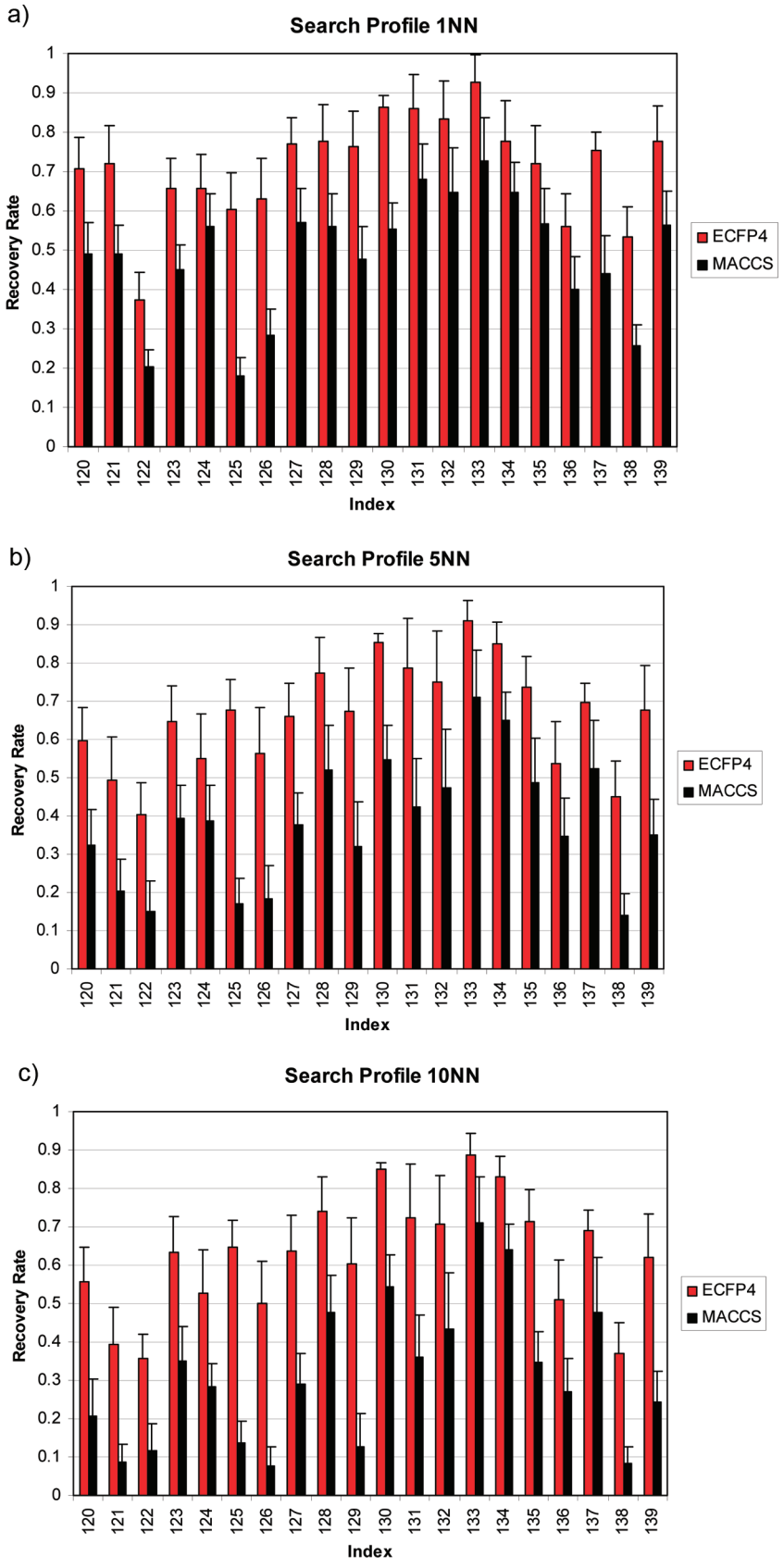


Figure 1. Similarity search profile. Average recovery rates (selection set size equal to the number of ADCs) of a representative subset of 20 activity classes (number 120−139 in Supporting Information Table S1) are reported in a histogram representation for MACCS (black) and ECFP4 (red). Positive standard deviations are displayed as error bars. The index on the *x*-axis reports the consecutively numbered activity classes. Search strategy: (a) 1NN, (b) 5NN, (c) 10NN.

As a well-established fingerprint that marks the basic performance level of conventional 2D fingerprints, we selected MACCS[23] as the prototype of a "low resolution" structural fragment dictionary fingerprint (consisting of 166 predefined structural keys). The MACCS design goes back to the roots of 2D fingerprinting and is often used as a standard to put the performance of different fingerprints into perspective.[10,16] Furthermore, we selected ECFP4 as a representative of a popular "high resolution" class of extended connectivity fingerprints[24] that currently probably represent the top performance level among 2D fingerprints of different design.[10,11] These combinatorial fingerprints systematically monitor circular atom environments up to a given bond diameter in test compounds and assemble these structural features in a molecule-specific manner, rather than based on predefined dictionaries. Hence, MACCS and ECFP4 can be used as markers to represent the current spectrum of 2D fingerprint search performance, which enables the generation of a similarity search profile of a large database and also makes it possible to characterize individual compound activity classes according to the degree of difficulty they represent for 2D fingerprint searching. Importantly, we did not aim to generate individual activity classes with predefined molecular properties for benchmarking or, alternatively, to carry out a standard fingerprint comparison. Rather, the focal points of this study have been to mark the boundaries of 2D similarity search performance on a large scale and, in addition, provide some guidance for the evaluation of similarity search calculations on well-curated publicly available compound classes.

Table 1. Average Recovery Rates[a]

	ECFP4			MACCS		
	1NN	5NN	10NN	1NN	5NN	10NN
average	63.6	58.9	55.2	45.3	37.2	31.7
st dev	7.4	8.6	8.2	7.2	8.4	7.6

[a] Average recovery rates (in percent) and standard deviations (st dev) are reported over all ChEMBL activity classes and search trials.

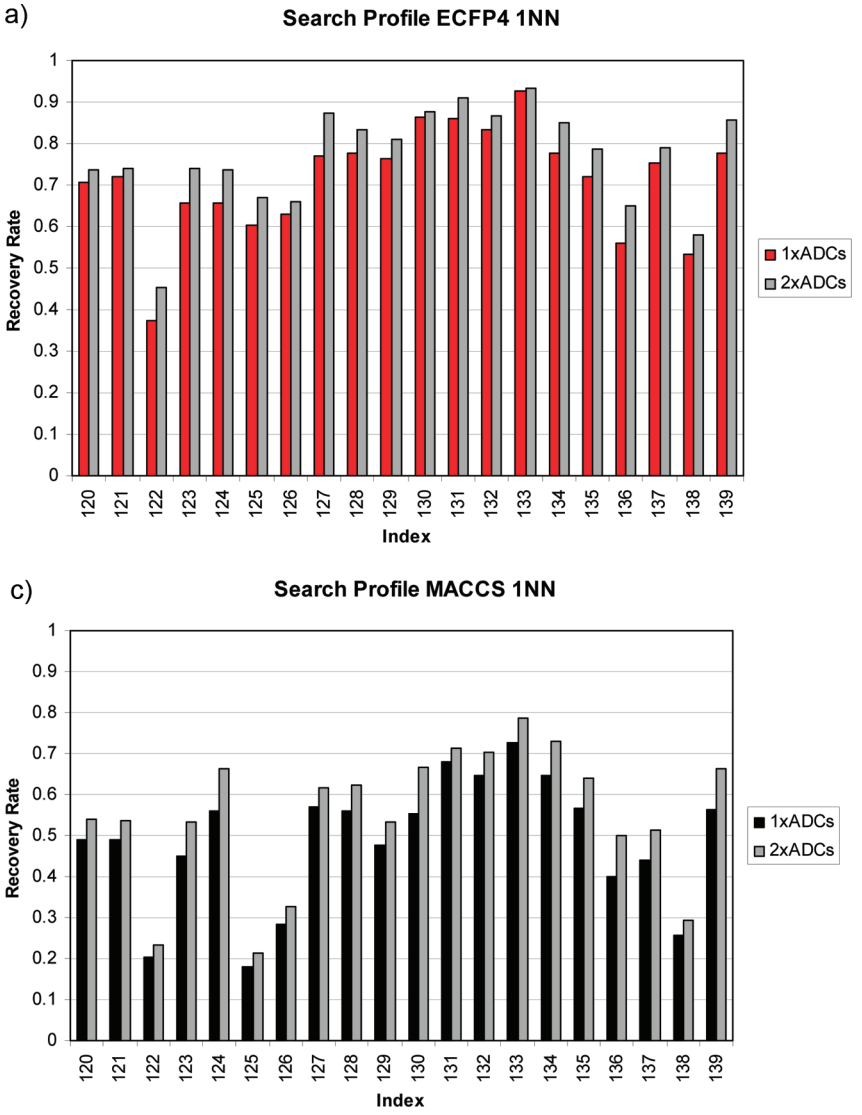


Figure 2. Early enrichment characteristics. Average recovery rates of a representative subset of 20 activity classes (numbers 120−139 in Supporting Information Table S1) are reported for selection set sizes of one or two times the number of ADCs per activity class. The index reports the consecutively numbered activity classes. Fingerprints and search strategies: (a) ECFP4/1NN, (b) MACCS/1NN.

Table 2. Activity Class Yielding Highest Fingerprint Search Performance[a]

			ECFP4		MACCS	
no.	target ID	target name	1NN	10NN	1NN	10NN
256	101174	pituitary adenylate cyclase-activating polypeptide type I receptor	100.0	100.0	100.0	100.0
264	101395	IgG receptor FcRn large subunit p51	100.0	99.2	100.0	100.0
83	10102	5-lipoxygenase activating protein	100.0	100.0	94.7	95.8
180	10144	bone morphogenetic protein 1	97.2	97.2	84.0	88.6
251	12909	ileal bile acid transporter	90.4	91.9	89.5	81.5
253	20130	inhibitor of apoptosis protein 3	90.3	90.1	86.0	86.6
169	275	retinoid X receptor alpha	92.5	94.9	79.2	84.0
228	11061	motilin receptor	94.3	90.7	83.3	82.0
173	10056	DNA-dependent protein kinase	88.8	93.5	81.6	84.5
231	11096	sodium/hydrogen exchanger 1	77.9	89.7	88.5	91.2
214	10845	phospholipase D1	90.4	90.6	81.1	84.3
246	11758	glucagon-like peptide receptor	95.8	80.3	87.7	78.4
189	11402	furin	91.3	81.1	91.0	76.7
31	12725	matriptase	91.4	83.6	87.9	76.0
262	101219	secreted frizzled-related protein 1	99.1	100.0	67.1	72.0
119	176	Purinergic receptor P2Y12	89.2	86.3	81.4	79.7
175	10087	deoxycytidine kinase	95.9	91.8	85.4	63.1
247	100098	serine/threonine-protein kinase WEE1	95.2	96.6	65.9	71.7
74	10624	serotonin 5a (5-HT5a) receptor	84.3	91.4	72.5	79.1
133	12659	prostanoid DP receptor	92.8	88.5	72.8	70.9
203	10582	cytosolic phospholipase A2	95.1	89.5	76.2	62.9
261	100862	metastin receptor	93.9	84.6	75.5	66.1
90	117	somatostatin receptor 2	87.4	86.4	71.1	73.8
6	4	voltage-gated T-type calcium channel alpha-1H subunit	86.5	80.4	74.6	76.3
216	11635	protein kinase C alpha	79.9	85.3	72.5	72.0
235	11242	Focal adhesion kinase 1	91.9	92.3	63.5	61.9
48	34	fibronectin receptor beta	93.2	90.3	68.0	56.9
212	100077	cell division cycle 7-related protein kinase	89.1	89.5	60.8	68.4
33	193	coagulation factor IX	85.6	86.0	60.9	74.7
102	80	FK506-binding protein 1A	91.0	85.6	71.0	59.0

[a] The top 30 activity classes yielding the highest overall search performance are reported and ranked according to decreasing average recovery rate (i.e., top-down) of ECFP4 (1NN and 10NN) and MACCS (1NN and 10NN) calculations for selection sets equal to the number of ADCs. For each activity class, the ChEMBL target ID and target name are provided and average recovery rates are reported (in percent).

The results of our large-scale fingerprint search investigation on ChEMBL are reported herein.

METHODS AND MATERIALS

Compound Data Sets. From ChEMBL, version 9,[22] activity classes were systematically extracted that contained at least 50 compounds active against human target proteins at high confidence level (ChEMBL level 9) for direct (D) interactions (i.e., 9/D[22]) with at least 10 μM potency. On the basis of these selection criteria, a total of 266 activity classes were obtained that contained between 50 and 1793 compounds, with on average ~239 compounds per class, as reported in Table S1 of the Supporting Information. These activity classes consist of designated enzyme or transporter inhibitors or receptor antagonists (with the exception of one class designated as ligands). When reporting activity classes herein, we refer to the target name, as given in ChEMBL.

Fingerprints and Search Strategies. For both MACCS[23] and ECFP4,[24] three *k*-nearest neighbor (*k*NN) search strategies[7] for multiple reference compounds were applied, i.e., 1NN, 5NN, and 10NN. The Tanimoto coefficient (Tc)[1] was calculated as the similarity measure. In 1NN calculations, a database compound is compared to all *k* reference compounds and the highest Tc value is utilized as the final similarity value for the database compound. In 5NN and 10NN calculations, the top 5 and top 10 Tc values are averaged, respectively, to yield the final similarity value for a database compound.

Similarity Searching. From each activity class, 100 reference sets of 10 compounds each were randomly selected and used for individual MACCS and ECFP4 search trials. In each case, all remaining active compounds were added as *active database compounds* (ADCs) to a background database containing one million molecules randomly selected from ZINC.[25] The choice of 10 reference compounds meant that the 10NN search strategy equally took similarity contributions from all reference compounds into account when calculating the similarity score for a database molecule. Initially, rather than using database selection sets of constant size, activity class-specific selection sets were utilized of a size equal to the number of ADCs. For each activity

Table 3. Activity Classes Yielding Lowest Fingerprint Search Performance[a]

			ECFP4		MACCS	
no.	target ID	target name	1NN	10NN	1NN	10NN
5	165	HERG	21.1	9.9	13.6	1.5
37	10193	carbonic anhydrase I	17.6	11.3	17.4	6.2
24	15	carbonic anhydrase II	17.6	14.3	15.8	7.0
96	11489	11-beta-hydroxysteroid dehydrogenase 1	25.1	16.8	15.2	2.8
67	121	serotonin transporter	26.5	17.5	14.7	5.5
62	72	dopamine D2 receptor	27.0	19.2	13.0	7.3
104	259	cannabinoid CB2 receptor	30.0	17.8	15.9	4.9
22	10188	MAP kinase p38 alpha	29.5	19.4	16.1	4.2
70	108	serotonin 2c (5-HT2c) receptor	30.7	19.8	18.7	5.4
34	12952	carbonic anhydrase IX	27.8	21.5	19.9	8.0
36	93	acetylcholinesterase	33.3	22.9	17.7	4.8
20	10980	vascular endothelial growth factor receptor 2	35.5	23.2	16.7	3.5
73	19905	melanin-concentrating hormone receptor 1	29.0	19.6	16.9	14.6
66	107	serotonin 2a (5-HT2a) receptor	35.2	20.2	20.9	4.6
103	87	cannabinoid CB1 receptor	31.8	31.3	14.9	9.9
91	17045	cytochrome P450 3A4	39.5	23.7	23.2	4.7
89	11140	dipeptidyl peptidase IV	36.9	27.3	17.8	9.5
117	114	adenosine A1 receptor	31.9	26.9	21.1	12.6
68	90	dopamine D4 receptor	33.4	20.6	24.3	14.2
255	100166	kinesin-like protein 1	40.4	28.4	19.0	6.9
26	13001	matrix metalloproteinase-2	33.1	27.6	24.9	9.4
92	104	monoamine oxidase B	38.5	24.8	21.5	10.6
40	65	cytochrome P450 19A1	31.8	30.1	21.3	12.4
55	61	muscarinic acetylcholine receptor M1	38.1	27.4	25.9	6.5
99	10280	histamine H3 receptor	36.6	29.5	20.3	12.2
53	51	serotonin 1a (5-HT1a) receptor	34.8	27.0	24.8	12.8
77	100	norepinephrine transporter	41.2	28.8	25.9	5.6
155	10260	vanilloid receptor	40.0	31.1	20.8	10.7
76	52	alpha-2a adrenergic receptor	40.5	28.3	28.0	6.8
153	11365	cytochrome P450 2D6	42.0	24.6	28.8	8.7

[a] The bottom 30 activity classes yielding the lowest overall search performance are reported and ranked according to increasing average recovery rate (i.e., bottom-up) of ECFP4 (1NN and 10NN) and MACCS (1NN and 10NN) calculations for selection sets equal to the number of ADCs. For each activity class, the ChEMBL target ID and target name are provided and average recovery rates are reported (in percent).

class, compound *recovery rates* (RRs) were then calculated by determining the ratio of active compounds contained in each class-specific selection set over all available ADCs. For example, if an activity class contained 200 compounds, 190 ADCs were available. If a search trial recovered 95 of these active compounds within a selection set of 190 database compounds (equal to the number of ADCs), the recovery rate would be 50%. Individual RRs were then averaged over all 100 trials for each activity class. Receiver operating characteristic (ROC) curves[26] and ROC area under the curve (AUC) values[26] were also calculated for averaged search trials. The initial use of ADC-based selection sets ensured that selection set sizes were balanced with respect to the size of an activity class and that in each case, a perfect similarity search outcome with 100% recovery rate (and 100% search specificity of the calculations) was principally possible. Subsequently, larger selection sets of two to three times the number of ADCs (e.g., 380 or 570 compounds for the example given above) were also considered. Finally, average RRs were also calculated for all classes for a constant database selection set size equal to the largest number of ADCs among all activity classes, i.e., 1783 compounds, which corresponded to ~0.18% of the screening database. MACCS and ECFP4 were generated using the Molecular Operating Environment[27] and Pipeline Pilot,[28] respectively, and all search calculations were carried out with in-house generated Java scripts.

RESULTS AND DISCUSSION

Similarity Search Profile. All 266 activity classes extracted from ChEMBL were subjected to systematic fingerprint search calculations in order to generate a similarity search profile of the database. From each class, 100 compound reference sets were randomly selected and for each combination of a fingerprint and a search strategy, 100 independent search trials were carried out in order to obtain statistically relevant data, which amounted to a total of ~160 000 search trials with multiple reference compounds. Figure S1 of the Supporting Information reports the resulting similarity search profiles for the three alternative nearest neighbor search strategies, and Figure 1 shows three representative profile subsets (for 20 activity classes, 120−139). In addition, Supporting Information Table S1 also reports 1NN recall

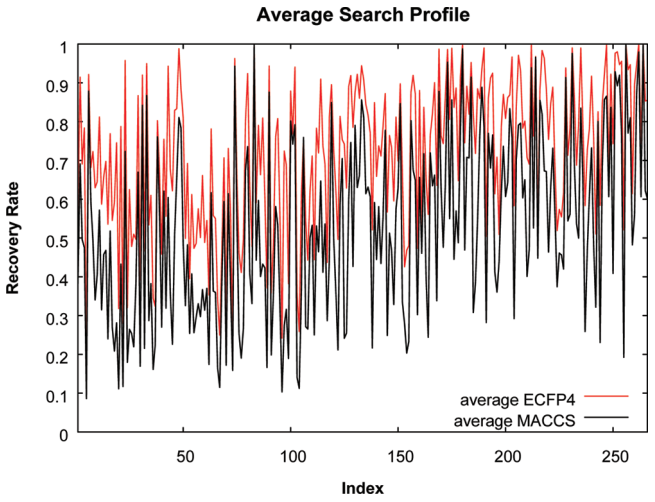


Figure 3. Similarity search profile for large database selection sets. For MACCS (black) and ECFP4 (red), recovery rates averaged over all three search strategies and for a constant database selection set size of 1783 molecules (see text) are plotted for all 266 activity classes. Index reports the consecutively numbered activity classes according to Supporting Information Table S1.

rates for each class. The profiles revealed the anticipated differences in global search performance between ECFP4 and MACCS. With only two exceptions, ECFP achieved consistently higher recovery rates (RRs). Furthermore, the profiles also illustrated the general compound class-dependence of fingerprint/similarity search calculations, with in part significantly varying RRs for each fingerprint. Importantly, however, the profiles revealed a perhaps unexpected success rate of 2D fingerprint searching on this large array of activity classes. In 203 of 266 test cases (~76%), an RR of at least 50% was obtained with at least the better performing fingerprint and at least one of the three different search strategies. It should be noted that these RRs were achieved for generally small selection set sizes equal to the number of ADCs for each class. For all search calculations, the average RR was 59.2% for ECFP4 and 38.1% for MACCS, which delineates a global performance range between approximately 40% and 60% compound recall achieved by a low-resolution (MACCS) and a high-resolution (ECFP4) 2D fingerprint. Given the large-scale character of these search calculations, these findings provide a realistic expectation value for 2D fingerprint searching on diverse compound classes.

Similarity Search Strategies. The 1NN, 5NN, and 10NN search strategies take contributions of reference compounds in different ways into account (see Methods and Materials). Because 1NN calculations only consider the match between a database molecule and the most similar reference compound, this strategy displays the tendency to select database molecules that are very similar to individual reference compounds. By contrast, because 10NN calculations take contributions of 10 reference compounds equally into account, this strategy shows a greater tendency to select database molecules that structurally differ from individual reference compounds. In Figure S2 of the Supporting Information, search profiles are compared for the 1NN, 5NN, and 10NN search strategies, and Table 1 reports the average RRs and standard deviations for each combination of a strategy and fingerprint. For both fingerprints, we observed that the global search performance decreased with increasing numbers of reference compound contributions (i.e., from 1NN over 5NN to 10NN), with an overall decline of ~8% for ECFP4 and 13% for MACCS. For all search calculations, standard deviations of ~7% or 8% were observed, which reflected the (limited) influence of reference set composition on the search results. Thus, we found that 1NN was the globally preferred nearest neighbor search strategy, yielding an average RR of 63.6% and 45.3% for ECFP4 and MACCS, respectively. Although the differences between individual search strategies were not very large, maximally on the order of 10%, selecting database molecules that were most similar to individual reference compounds globally produced highest RRs on the ChEMBL activity classes. These findings were consistent with the notion that compound data sets from medicinal chemistry typically contain different series of analogs, which are often easier to detect when applying the 1NN rather than other *k*NN search strategies.

Enrichment Behavior. We also studied the enrichment characteristics in database selection sets of increasing size. Figure S3 of the Supporting Information shows similarity search profiles for the original selection set sizes and selection sets that were doubled in size, and Figure 2 shows representative profile subsets for 20 activity classes and two fingerprint/search strategy combinations. Profile subsets for the remaining four fingerprint/strategy combinations are shown in Figure S4 of the Supporting Information. For both fingerprints, we consistently observed only slight increases in RRs of a few percent when selection sets were doubled or tripled in size (data not shown). Thus, these fingerprint search calculations were generally characterized by an early enrichment of active compounds in database selection sets. This meant that correctly identified active compounds often appeared at relatively high positions in the Tc-based similarity rankings. These finding were also consistent with the observation that active compounds were preferentially detected by matching the most similar reference compound (1NN), which typically yields higher similarity values than Tc average calculations and hence increases the probability of higher ranking positions. Figure S5 of the Supporting Information shows representative ROC curves for ECFP4 and MACCS 1NN calculations at different levels of search performance and ROC AUC values for all activity classes are reported in Table S2 of the Supporting Information.

Prioritization of Activity Classes. Our low/high resolution fingerprint similarity search strategy also made it possible to categorize activity classes according to their relevance for fingerprint benchmarking. We first identified particularly "easy" and "difficult" classes for 2D fingerprinting. In Table 2, the top 30 classes with overall highest search performance are reported. For ECFP4, the search performance was consistently very high in these cases, at or above the 90% levels, for both 1NN and 10NN calculations. For MACCS, RRs of close to or above 80% were also observed for 16 classes and all remaining RRs were above 60%. For the first three classes, almost perfect search results were obtained for both ECFP4 and MACCS. Taken together, the activity classes listed in Table 2 consistently yielded high to very high search performance for our prototypic low- and high-resolution 2D fingerprints. Thus, these classes are not suitable for fingerprint benchmarking because they yield RRs that go much beyond the typical performance range of 2D fingerprints, even for relatively small database selection sets. Importantly, the classes in Table 2 include a number of popular targets, for example, phospholipases, serine proteases, protein kinases, purinergic receptors, and other G protein coupled receptors that might often be attractive for benchmark trials. However, the uncritical choice of such data sets would provide artificially good results for fingerprint methods.

Table 4. Activity Classes Preferred for Evaluating 2D Fingerprints[a]

					average RR	
no.	target ID	target name	BMS	cpds per CSK	ECFP4	MACCS
4	11359	phosphodiesterase 4D	60	3.30	78.4	47.5
8	28	thymidylate synthase	44	4.29	72.3	49.4
9	11536	ghrelin receptor	228	3.52	63.0	34.1
10	8	tyrosine-protein kinase ABL	64	4.47	64.5	40.5
12	10434	tyrosine-protein kinase SRC	229	3.48	58.7	31.5
13	12670	tyrosine-protein kinase receptor FLT3	49	3.30	65.7	45.7
14	20014	serine/threonine-protein kinase Aurora-A	66	3.65	69.7	46.8
16	234	insulin-like growth factor I receptor	124	4.09	76.9	51.8
21	12261	c-Jun N-terminal kinase 1	51	5.94	78.7	43.3
35	12209	carbonic anhydrase XII	60	3.40	61.0	38.2
42	25	glucocorticoid receptor	169	4.41	55.6	31.9
44	36	progesterone receptor	99	5.79	67.9	36.2
52	43	beta-2 adrenergic receptor	88	2.11	69.2	48.2
54	219	muscarinic acetylcholine receptor M3	140	2.60	61.2	40.8
57	130	dopamine D3 receptor	214	3.23	57.5	33.0
59	105	serotonin 1d (5-HT1d) receptor	45	1.81	66.1	36.7
81	11336	neuropeptide Y receptor type 5	182	6.33	63.3	40.8
86	20174	G protein-coupled receptor 44	132	5.21	66.3	40.1
95	126	cyclooxygenase-2	117	5.92	56.0	32.7
98	11225	renin	183	5.34	68.8	31.6
105	12252	beta-secretase 1	246	3.31	61.7	37.3
112	11682	glycine transporter 1	66	3.95	78.3	53.1
113	134	vasopressin V1a receptor	110	2.54	72.0	46.5
115	116	oxytocin receptor	55	4.03	73.7	41.2
120	11265	somatostatin receptor 5	67	2.50	73.0	50.0
121	10475	neuropeptide Y receptor type 1	66	4.70	62.8	36.9
129	12679	C5a anaphylatoxin chemotactic receptor	67	3.54	78.6	42.7
140	10579	C−C chemokine receptor type 4	87	2.73	65.9	44.5
142	11575	C−C chemokine receptor type 2	178	6.11	71.2	43.5
143	18061	sodium channel protein type IX alpha subunit	58	5.26	78.8	55.3
146	237	leukotriene A4 hydrolase	87	3.20	76.4	51.3
147	276	phosphodiesterase 4A	38	2.61	73.3	46.7
148	11534	cathepsin S	298	3.61	59.6	32.9
152	10198	voltage-gated potassium channel subunit Kv1.5	97	3.94	67.0	33.6
163	10498	cathepsin L	67	3.29	65.7	40.2
168	12911	cytochrome P450 2C9	31	2.27	63.6	33.8
171	12968	orexin receptor 2	43	4.55	74.3	47.6
181	100579	nicotinic acid receptor 1	80	4.47	74.6	46.9
186	100126	serine/threonine-protein kinase B-raf	73	2.94	71.8	38.7
195	10378	cathepsin B	56	2.56	61.4	41.3
196	10417	P2X purinoceptor 7	69	3.26	70.9	36.1
210	10752	inhibitor of nuclear factor kappa B kinase beta subunit	46	3.81	70.8	40.1
211	10773	interleukin-8 receptor B	76	6.85	69.0	47.1
213	11631	sphingosine 1-phosphate receptor Edg-1	51	3.59	76.3	52.6
220	10927	urotensin II receptor	74	3.00	75.4	46.4
230	11085	melatonin receptor 1B	52	3.61	78.4	56.2
234	11442	liver glycogen phosphorylase	104	5.10	79.3	50.0
238	11279	metabotropic glutamate receptor 1	84	4.37	72.6	46.7
241	11488	estradiol 17-beta-dehydrogenase 3	39	5.30	76.3	48.4
250	12840	macrophage colony stimulating factor receptor	59	5.57	74.3	40.9

[a] Listed are 50 activity classes that met our selection criteria for benchmarking relevance (as described in the text). These classes are ordered by their consecutive numbers. Average RRs over all search strategies are reported (in percent) for ECFP4 and MACCS and a database selection set size of 1783 compounds (see text for details). For each class, the total number of Bemis and Murcko scaffolds (BMS)[29] and the compound-to-carbon skeleton (CSK)[30] ratio (cpds per CSK) are reported.

dx.doi.org/10.1021/ci200199u |J. Chem. Inf. Model. 2011, 51, 1831–1839

In Table 3, we report the opposite end of the similarity search spectrum. Here the 30 activity classes with lowest search performance are listed. For ECFP4, these classes mostly resulted in RRs of ~20% to ~30%. For MACCS, many 10NN RRs were lower than 10% or even 5%, but 1NN RRs were still close to or above 20% in many instances. Therefore, fingerprint searching on none of these classes could *per se* be considered a failure. However, given their overall low search performance, the 2D fingerprints clearly approached their detection limits in these cases that also included a number of popular enzyme and G protein coupled receptor targets. Hence, these activity classes might be more appropriate for the evaluation of similarity methods that employ more elaborate molecular representations or utilize 3D information.

In virtual screening benchmark calculations, compound recall is typically evaluated on the basis of larger selection set sizes than the variably balanced selection set sizes that we utilized for our analysis up to this point. Often, 0.1−1% of the screening/background database are selected for recovery rate analysis. Therefore, we also calculated average RRs over all search strategies for a selection set of constant size, i.e., the largest individual selection set utilized in our study, which contained 1783 compounds corresponding to ~0.18% of our background database. On average, this constant selection set corresponded to an approximately 6-fold increase in selection set size for the ChEMBL activity classes. The resulting similarity search profile for all 266 classes is displayed in Figure 3. As expected, for this comparably large selection set, average RRs were higher than originally observed, with 71.9% and 52.7% for ECFP4 and MACCS, respectively (again with standard deviations of ~8%). However, the increase relative to the originally observed RRs was also limited with approximately 11% for ECFP4 and 14% for MACCS, consistent with the generally observed early enrichment characteristics.

On the basis of these results, we then prioritized activity classes that were considered particularly suitable for benchmarking of 2D fingerprints. Therefore, in light of the observed search performance range for our fingerprint prototypes, we selected activity classes that minimally yielded more than 30% compound recall for MACCS (thus ensuring a meaningful base performance) and maximally less than 80% recall for ECFP4 (thus leaving room for further improvements) and that differed by more than 20% in relative search performance (thus reflecting the overall performance range). On the basis of these selection criteria, we identified a total of 50 activity classes that we would assign a high priority for the evaluation and comparison of alternative 2D fingerprints. As reported in Table 4, these classes covered a variety of different target families including a number of prominent therapeutic targets and were generally characterized by the presence of large numbers of distinct scaffolds and low compound-to-carbon skeleton ratios (i.e., structural heterogeneity). These activity classes can also be obtained via the following URL (please, see the "Downloads" section): http://www.lifescienceinformatics.uni-bonn.de.

CONCLUSIONS

Here we have reported a large-scale similarity search investigation to systematically analyze compound activity classes extracted from the ChEMBL database, a major public domain repository of compounds originating from medicinal chemistry sources. For similarity search profiling of ChEMBL, we selected two prototypic 2D fingerprints that represent markers for current performance levels of popular fingerprints. On the basis of systematic search calculations, we also determined the global performance range defined by these fingerprints covering compound data sets directed against the spectrum of current pharmaceutical targets. Overall, the search results were rather encouraging, more so than we anticipated, indicating that many activity classes can be well treated using 2D fingerprints, despite their relative simplicity. Other general trends emerged concerning preferred search strategies and early enrichment characteristics of active compounds that corroborated earlier findings. Furthermore, by comparing the search performance of our low- and high-resolution fingerprint standards, we have identified activity classes that were unsuitable for 2D fingerprint evaluation, because they yielded artificially high search performance, and other classes that represented rather difficult test cases where 2D fingerprints approached their limits. We also prioritized 50 activity classes as particularly useful for 2D fingerprint evaluation in light of the search characteristics we observed. Taken together, these findings should also aid in the design of meaningful benchmark investigations.

ASSOCIATED CONTENT

Supporting Information. Supplementary Table S1 listing all activity classes extracted from the ChEMBL database and reports recovery rates for each class, Supplementary Table S2 reporting ROC AUC values, and Supplementary Figures S1−S5 showing similarity search profiles, comparisons of similarity search strategies, enrichment characteristics for all activity classes, enrichment characteristics for activity class subsets, and exemplary ROC curves for activity classes at different similarity search performance levels, respectively. This information is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Tel.: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

REFERENCES

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

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

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

(6) Nisius, B.; Bajorath, J. Reduction and recombination of fingerprints of different design increase compound recall and the structural diversity of hits. *Chem. Biol. Drug Des.* **2010**, *75*, 152–160.

(7) 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.

(8) 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.

(9) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. New Methods for Ligand-Based Virtual Screening:

Use of Data Fusion and Machine Learning to Enhance the Effectiveness of Similarity Searching. *J. Chem. Inf. Model* **2006**, *46*, 462–470.

(10) 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.

(11) 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.

(12) 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.

(13) 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, 2011; pp 73−103.

(14) 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.

(15) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Topological Descriptors for Similarity-Based Virtual Screening Using Multiple Bioactive Reference Structures. *Org. Biomol. Chem.* **2004**, *2*, 3256–3266.

(16) Sastry, M.; Lowrie, J. F.; Dixon, S. L.; Sherman, W. Large-Scale Systematic Analysis of 2D Fingerprint Methods to Improve Virtual Screening Enrichments. *J. Chem. Inf. Model* **2010**, *50*, 771–784.

(17) *MDL Drug Data Report*; Accelrys: San Diego, CA, 2011.

(18) Rohrer, S. G.; Baumann, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J. Chem. Inf. Model* **2009**, *49*, 169–184.

(19) Tiikkainen, P.; Markt, P.; Wolber, G.; Kirchmair, J.; Distinto, S.; Poso, A.; Kallioniemi, O. Critical Comparison of Virtual Screening Methods Against the MUV Data Set. *J. Chem. Inf. Model* **2009**, *49*, 2168–2178.

(20) *PubChem*; National Center for Biotechnology Information: Bethesda, MD, 2010; http://pubchem.ncbi.nlm.nih.gov/ (accessed December 10, 2010).

(21) 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.

(22) *ChEMBL*; European Bioinformatics Institute (EBI): Cambridge, 2011; http://www.ebi.ac.uk/chembl/ (accessed March 2, 2011).

(23) *MACCS Structural keys*; Accelrys: San Diego, CA, 2011.

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

(25) 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.

(26) Bradley, A. P. The Use of the Area under the ROC Curve for the Evaluation of Machine Learning Algorithms. *Pattern Recog.* **1997**, *30*, 1145–1159.

(27) *Molecular Operating Environment*; Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2010.

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

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

(30) 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.