J Comput Aided Mol Des (2012) 26:1101–1109
DOI 10.1007/s10822-012-9603-9

Systematic assessment of scaffold distances in ChEMBL: prioritization of compound data sets for scaffold hopping analysis in virtual screening

Ruifang Li · Jürgen Bajorath

Received: 18 July 2012 / Accepted: 7 September 2012 / Published online: 13 September 2012

Abstract The evaluation of the scaffold hopping potential of computational methods is of high relevance for virtual screening. For benchmark calculations, classes of known active compounds are utilized. Ideally, such classes should have a well-defined content of structurally diverse scaffolds. However, in reported benchmark investigations, the choice of activity classes is often difficult to rationalize. To provide a compendium of well-characterized test cases for the assessment of scaffold hopping potential, structural distances between scaffolds were systematically calculated for compound classes available in the ChEMBL database. Nearly seven million scaffold pairs were evaluated. On the basis of the global scaffold distance distribution, a threshold value for large scaffold distances was determined. Compound data sets were ranked based on the proportion of scaffold pairs with large distances they contained, taking additional criteria into account that are relevant for virtual screening. A set of 50 activity classes is provided that represent attractive test cases for scaffold hopping analysis and benchmark calculations.

Keywords Scaffolds · Structural distances · Database analysis · Scaffold hopping · Ligand-based virtual screening

Introduction

Scaffold hopping [1–5] is the ultimate goal of ligand-based virtual screening (LBVS) [5–7] and the most important criterion for prospective LBVS applications [7]. Accordingly, the assessment of scaffold hoping potential of LBVS methods and calculation protocols is a major aspect of benchmark investigations [5, 6].

Scaffold hopping generally refers to the ability of an LBVS method to identify novel active compounds with core structures (scaffolds, frameworks) that are distinct from those of reference molecules. The assessment of scaffold hopping potential is complicated by the fact that scaffolds can be defined and compared in different ways [8]. Moreover, formally distinct scaffolds might span a wide range of structural relationships [8], ranging from chemically very similar scaffolds (e.g., core structures only distinguished by a heteroatom substitution) to virtually unrelated ones. Thus, the degree of difficulty involved in detecting structurally diverse active compounds with different scaffolds is not accounted for by statistical analysis of scaffold hops [6] because the magnitude of structural differences between reference molecules and correctly identified hits is not taken into consideration.

In order to address this critical issue for LBVS method evaluation, we have recently introduced a generally applicable method to calculate the structural distance between scaffolds, regardless of their chemical composition, topology, or size [9]. Applying this scaffold distance function, chemically intuitive structural distances have been obtained for a variety of scaffolds [9]. Hence, the distance function makes it possible to quantify the degree of difficulty involved in scaffold hopping exercises, i.e., the larger the distance is, the more difficult it is to detect two scaffolds sharing similar activity. Hence, applying this approach, it is possible to evaluate the outcome of LBVS trials in quantitative terms, beyond compound recall statistics.

For benchmark evaluation and comparison of LBVS methods, the selection of compound activity classes is

R. Li · J. Bajorath (✉)
Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität Bonn, Dahlmannstr. 2, 53113 Bonn, Germany
e-mail: bajorath@bit.uni-bonn.de

another critical aspect. It is generally difficult to rationalize the choice of activity classes for scaffold hopping calculations. Attempts have been made to design benchmark sets that exclusively consist of reference and active database compounds with distinct scaffolds such that every correctly identified active compound represents a scaffold hop [5]. However, the generation of such compound sets is time-consuming and laborious and still requires the calculation of scaffold distances to estimate the degree of difficulty.

In this study, we have attempted to apply the scaffold distance approach to systematically search for activity classes that provide challenging test cases for the assessment of scaffold hopping potential. For this purpose, we have calculated, analyzed, and compared intra-class scaffold distances for all activity classes currently available in ChEMBL [10]. Activity classes have been ranked on the basis of scaffold distance information and characterized taking additional criteria into account that are relevant for LBVS. On the basis of distance-based ranking and the additional information provided, compound classes can be selected from ChEMBL that present meaningful test cases for LBVS benchmark investigations.

Methods

Compound data sets

From ChEMBL, version 13 [10, 11], activity classes were systematically selected that contained compounds active against human target proteins at high confidence level (ChEMBL level 9) for direct (D) interactions [11]. In addition, we required that active compounds had a reported potency (K_i or IC_{50} values) of at least 10 μM (to avoid the selection of very weakly or nearly inactive compounds). For compounds comprising such activity classes, it can a priori not be assumed that all compounds act by the same molecular mechanism. This might often even be unlikely. Scaffolds were defined according to Bemis and Murcko (BM) [12] by removing all R-groups from ring-containing compounds. In order to avoid artificial scaffold hopping scenarios, we omitted BM scaffolds and corresponding compounds from all data sets that contained large aliphatic ring structures or cyclic peptide substructures with more than eight ring atoms. Such macrocycles often produce large structural distances compared to drug-like scaffolds (as expected) and are not relevant for scaffold hopping analysis in the context of LBVS. We extracted all compound activity classes from ChEMBL that met the selection criteria specified above and yielded at least three distinct BM scaffolds. A total of 790 compound activity classes were obtained. The number of compounds and BM scaffolds in these classes ranged from 3 to 2,212 and 3 to 931, respectively (with an average of ~71 scaffolds per class).

Scaffold distance method

The scaffold distance method applied herein has been described in detail [9]. Briefly, the methodology involves the following steps. Initially, a pair of scaffolds is rendered topologically equivalent through defined molecular editing procedures. For topologically equivalent pairs, ring correspondence is determined. Then, ring and linker substructures comprising each scaffold are transformed into 1D atom sequences. All possible alignments between ring and linker sequences are generated and scored on the basis of an atom type similarity matrix (yielding normalized similarity values). The best-scoring alignment of scaffold atom sequences is selected and its similarity is converted into a scaffold distance D = 1-similarity. Hence, scaffold distances range from 0 to 1. Figure 1 shows exemplary series of scaffolds with defined chemical changes and the results of pair-wise scaffold distance comparisons. As can be seen, calculated distances are chemically intuitive. Scaffold distance calculations are based on molecular graphs. Hence, stereochemical differences between scaffolds are not taken into account.

Distance calculations

For each qualifying compound class, pair-wise distances between all scaffolds were systematically calculated. From the global distribution of observed scaffold distances, a threshold value for large scaffold distances was determined as two times the standard deviation above the mean. Scaffold pairs exceeding this threshold value were designated as 'Large scaffold Distance' (LD) pairs. In addition, LD pairs with limited size differences of no more than 10 non-hydrogen (Heavy) atoms were identified (LD_H pairs). Systematic extraction of BM scaffolds from ChEMBL compound sets and all scaffold distance calculations were carried out with in-house generated programs.

Similarity search calculations

For highly ranked activity classes, similarity search control calculations were carried out using combinations of two fingerprints, MACCS structural keys [13] and Extended Connectivity Fingerprint with Bond Diameter 4 (ECFP4) [14], with two nearest neighbor (NN) search strategies, 1-NN and 10-NN. In 1-NN calculations, the largest Tanimoto coefficient (Tc) value obtained relative to a reference compound is used as the final similarity score of a database compound. In 10-NN calculations, Tc values are averaged

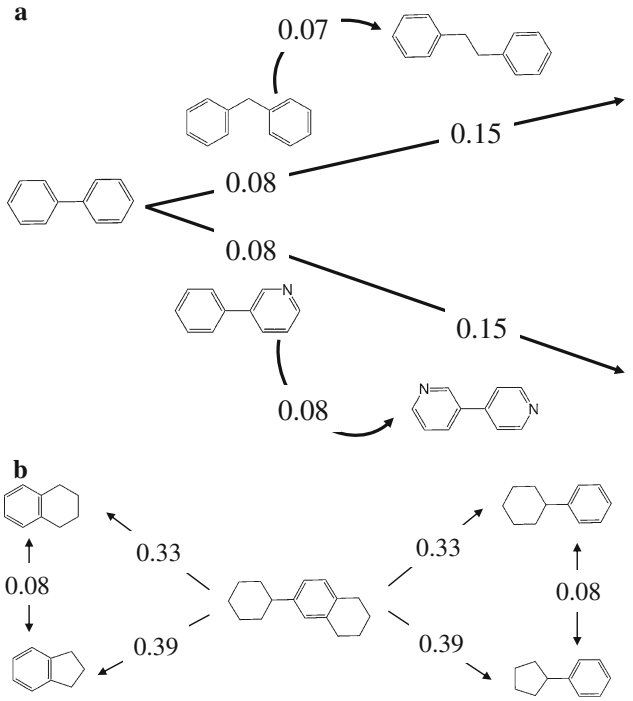

Fig. 1 Scaffold distances. Results of exemplary scaffold distance calculations are reported. Compared scaffolds are systematically distinguished by different chemical modifications and include series of scaffolds with **a** the same topology and ring size but varying linker length and heteroatom content and **b** different topology and defined differences in ring size and/or content. For each pair-wise comparison, the scaffold distance is reported

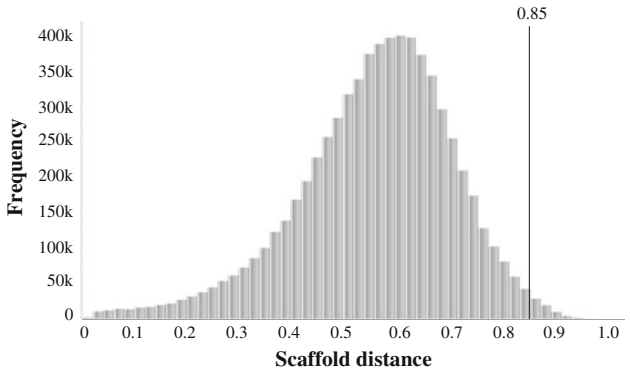

Fig. 2 Global scaffold distance distribution. Shown is the global distribution of scaffold distances for 6,793,494 unique pairs of scaffolds from 790 screening activity classes. The LD threshold value is marked

for 10 reference compounds and the resulting value is assigned as the final similarity score to a database compound. For each activity class, 100 independent search trials were carried out for each combination with 100 sets of 10 randomly selected reference compounds. In each case, the remaining active compounds were added as potential hits to a database of one million compounds randomly selected from ZINC [15]. Hit and recovery rates were calculated for database selection sets of 400 compounds and averaged over 100 search trials.

Results and discussion

Scaffold hopping evaluation

The major goal of our study has been to rationalize the selection of compound activity classes for the evaluation of the scaffold hopping potential of LBVS methods. For this purpose, we have systematically analyzed scaffold distances in ChEMBL compound data sets and identified activity classes that contained a significant proportion of scaffold pairs with large scaffold distances.

Global scaffold distance distribution

For all 790 activity classes, intra-class scaffold distances were systematically calculated. Then, the global distribution of distances was determined over all classes using all 6,793,494 possible scaffold pairs, as shown in Fig. 2. The bell-shaped distribution spanned nearly the entire range of possible distances, with a mean of 0.554 (corresponding to the average scaffold distance in ChEMBL) and a standard deviation of 0.148. The threshold value for LD pairs was calculated as 0.850 (two standard deviations above the mean). These LD pairs represented a small subset of all scaffold pairs with the largest distances (Fig. 2). With increasing numbers of LD pairs within a class, scaffold hopping challenges also increase.

Characterization of activity classes

We next characterized the scaffold information contained in all 790 classes in detail. The number of compounds and scaffolds per class, the compound-to-scaffold ratio, and the average scaffold distance were calculated. In addition, the number of LD pairs, the number of unique scaffolds they contained, and the proportion of LD pairs among all scaffold pairs within a class were determined. Because significant size differences between scaffolds often complicate scaffold hopping calculations, the number of LD pairs with limited size difference (of no more than 10 non-hydrogen atoms) and the number of unique scaffolds this subset of LD pairs contained was also determined.

Ranking of activity classes

The scaffold information for activity classes can be compared in different ways. Following our scaffold

Table 1 Activity classes for scaffold hopping evaluation

No	TID	Target name	No Cpds	No BMS	Cpds/ BMS	Avg ScDist	SD ScDist	LDP	LDP BMS	%LDP	LDP_H	LDP_H BMS
1	184	Glutamate carboxypeptidase II	80	19	4.2	0.627	0.287	40	19	23.4	23	15
2	10258	T cell protein-tyrosine phosphatase	58	30	1.9	0.662	0.209	86	22	19.8	68	18
3	12425	Nitric oxide synthase, inducible	252	119	2.1	0.631	0.218	990	109	14.1	61	24
4	12923	Carbonic anhydrase XIII	26	21	1.2	0.671	0.201	26	15	12.4	10	11
5	120	Steroid 5-alpha-reductase 2	64	31	2.1	0.529	0.268	52	19	11.2	49	18
6	11165	3-beta-hydroxysteroid-delta(8),delta(7)-isomerase	36	31	1.2	0.632	0.175	49	19	10.5	35	15
7	12054	Arachidonate 15-lipoxygenase	37	24	1.5	0.654	0.177	29	19	10.5	12	8
8	13061	Protein-tyrosine phosphatase 1B	336	153	2.2	0.617	0.179	1,149	153	9.9	577	88
9	10186	Carbonic anhydrase VI	90	49	1.8	0.624	0.182	103	40	8.8	38	26
10	11574	Interleukin-8 receptor A	74	26	2.8	0.466	0.220	26	19	8.0	23	17
11	12247	Protein kinase C delta	151	74	2.0	0.539	0.223	211	71	7.8	26	25
12	11398	P-glycoprotein 1	57	41	1.4	0.562	0.218	59	24	7.2	57	24
13	11902	Nerve growth factor receptor Trk-A	62	28	2.2	0.591	0.219	26	19	6.9	25	18
14	10553	Voltage-gated potassium channel subunit Kv1.3	124	59	2.1	0.375	0.277	111	42	6.5	111	42
15	12896	Carbonic anhydrase IV	100	51	2.0	0.615	0.177	82	43	6.4	26	20
16	11060	Carbonic anhydrase VII	154	67	2.3	0.598	0.193	141	53	6.4	46	31
17	11016	Carbonic anhydrase VB	126	58	2.2	0.587	0.198	105	45	6.4	38	28
18	11042	Carbonic anhydrase VA	144	65	2.2	0.604	0.189	130	52	6.3	41	29
19	20127	11-beta-hydroxysteroid dehydrogenase 2	73	33	2.2	0.657	0.193	31	13	5.9	27	12
20	11232	Histone deacetylase 4	59	37	1.6	0.588	0.185	39	27	5.9	18	18
21	11063	Carbonic anhydrase XIV	224	78	2.9	0.606	0.175	162	60	5.4	42	26
22	42	Trypsin I	74	42	1.8	0.526	0.205	46	26	5.3	21	22
23	11409	Dual specificity mitogen-activated protein kinase kinase 1	129	39	3.3	0.534	0.203	38	31	5.1	15	16
24	235	Leukocyte elastase	309	137	2.3	0.603	0.168	471	135	5.1	13	15
25	11635	Protein kinase C alpha	235	85	2.8	0.606	0.186	180	71	5.0	36	35
26	12622	Telomerase reverse transcriptase	134	56	2.4	0.551	0.234	76	56	4.9	30	31
27	11631	Sphingosine 1-phosphate receptor Edg-1	197	90	2.2	0.562	0.175	173	77	4.3	39	40
28	12209	Carbonic anhydrase XII	483	165	3.0	0.577	0.182	559	141	4.1	100	66
29	10083	Serine/threonine protein phosphatase PP1-alpha catalytic subunit	62	44	1.4	0.519	0.166	39	27	4.1	29	23
30	10880	Protein kinase C gamma	61	32	2.0	0.556	0.242	20	17	4.0	17	17
31	15	Carbonic anhydrase II	1,081	383	2.8	0.591	0.166	2,848	377	3.9	840	286
32	12952	Carbonic anhydrase IX	794	308	2.6	0.599	0.160	1,750	300	3.7	387	222
33	10878	Protein kinase C epsilon	73	32	2.3	0.559	0.222	18	15	3.6	15	15
34	11291	Anandamide amidohydrolase	352	127	2.8	0.568	0.163	286	121	3.6	15	18
35	93	Acetylcholinesterase	743	411	1.8	0.614	0.152	2,960	407	3.5	407	100
36	19	Estrogen receptor alpha	695	249	2.8	0.581	0.164	1,061	248	3.4	228	70
37	13053	Estradiol 17-beta-dehydrogenase 2	125	45	2.8	0.395	0.196	34	35	3.4	34	35
38	10938	Tyrosine-protein kinase JAK2	167	111	1.5	0.577	0.149	206	100	3.4	50	48
39	103719	Polyadenylate-binding protein 1	77	65	1.2	0.587	0.159	67	50	3.2	41	41
40	10809	Sphingosine 1-phosphate receptor Edg-6	251	122	2.1	0.560	0.169	230	108	3.1	51	53
41	12214	Tyrosine-protein kinase ZAP-70	80	32	2.5	0.448	0.207	15	16	3.0	12	13
42	11638	MAP kinase ERK2	54	31	1.7	0.522	0.198	14	15	3.0	10	11

Table 1 continued

No	TID	Target name	No Cpds	No BMS	Cpds/ BMS	Avg ScDist	SD ScDist	LDP	LDP BMS	%LDP	LDP_H	LDP_H BMS
43	11636	Protein kinase C beta	96	40	2.4	0.559	0.234	23	20	2.9	20	20
44	136	Delta opioid receptor	1,174	492	2.4	0.636	0.152	3,542	422	2.9	741	189
45	10193	Carbonic anhydrase I	925	345	2.7	0.584	0.164	1,700	336	2.9	419	252
46	10532	Butyrylcholinesterase	545	330	1.7	0.588	0.172	1,481	315	2.7	312	88
47	56	Androgen receptor	633	166	3.8	0.626	0.162	363	100	2.7	98	42
48	137	Kappa opioid receptor	1,578	651	2.4	0.637	0.144	5,569	600	2.6	1,581	292
49	65	Cytochrome P450 19A1	540	198	2.7	0.524	0.166	483	75	2.5	435	70
50	101400	Smoothened homolog	140	58	2.4	0.431	0.210	37	20	2.2	37	20

Listed are the top 50 compound classes from ChEMBL ranked by decreasing percentages of scaffold pairs exceeding the empirical large scaffold distance threshold value (0.850). For each class, the ChEMBL target ID (TID) and target name (Target name) are provided. In addition, the number of compounds per class (No Cpds), number of BM scaffolds (No BMS), and the compound-to-scaffold ratio (Cpds/BMS) are reported. All pair-wise scaffold distances within a class were calculated and the average (Avg ScDist) and standard deviation (SD ScDist) are provided. Furthermore, the number of scaffold pairs exceeding the large scaffold distance threshold value (LD Pairs, LDP), their percentage of all possible pairs (%LDP), and the number of unique scaffolds forming these pairs (LDP BMS) are reported. For LDP, the subset of scaffold pairs with a difference of max. 10 non-hydrogen atoms is given (LDP_H) as well as the number of scaffold forming these pairs (LDP_H BMS)

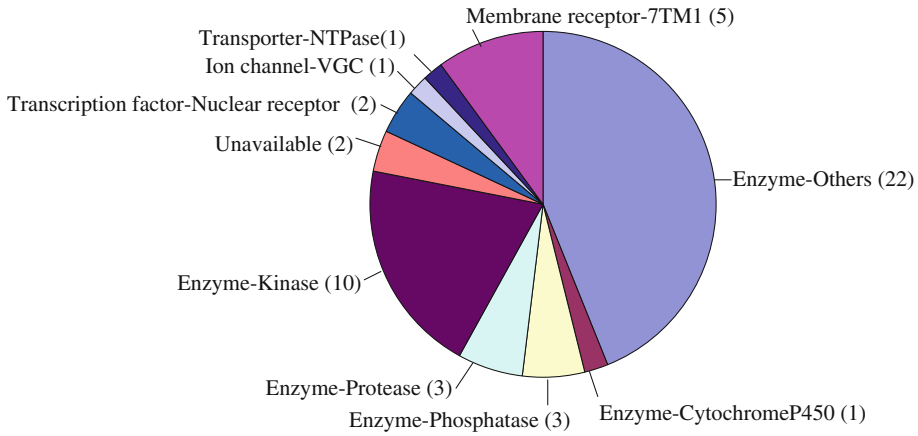

Fig. 3 Targets. For the top 50 activity classes according to Table 1, the target and family distribution is reported following the ChEMBL target classification scheme. The number of targets falling into each family is given in *parentheses*

distance-based approach, we found a comparison of different classes on the basis of their LD pairs most informative. Table 1 reports the top 50 activity classes from ChEMBL ranked by the proportion of LD pairs among all scaffold pairs within a class (%LDP). The maximum proportion of LD per class detected in ChEMBL was 23.4 % for glutamate carboxypeptidase II inhibitors (rank 1 in Table 1), followed by 19.8 % for T cell protein-tyrosine phosphatase inhibitors. The first seven activity classes contained more than 10 % LD pairs. The top 50 classes contained varying numbers of compounds and scaffolds, ranging from 26 to 1,578 and 19 to 651, respectively. A maximum compound-to-scaffold ratio of 4.2 was observed (rank 1), followed by 3.8 (rank 47). However, most ratios were around 2. Hence, on the basis of compound-to-scaffold ratios, the classes were structurally diverse. Furthermore, as illustrated in Fig. 3, the activity classes covered a wide range of target families, mostly different types of enzymes and membrane receptors. However, 11 classes consisted of inhibitors of various isoforms of carbonic anhydrase, which represented the largest family. Table 1 clearly shows that average scaffold distances were not a sensitive measure to prioritize classes, as we anticipated. Regardless of the size and composition of activity classes, average scaffold distances mostly fluctuated around 0.6 (with a few exceptions), i.e., relatively close to the average scaffold distance in ChEMBL (Fig. 2). Among the ranked classes, the largest average distance was 0.671 (rank 4) and the smallest 0.375 (rank 14). Standard deviations generally were on the

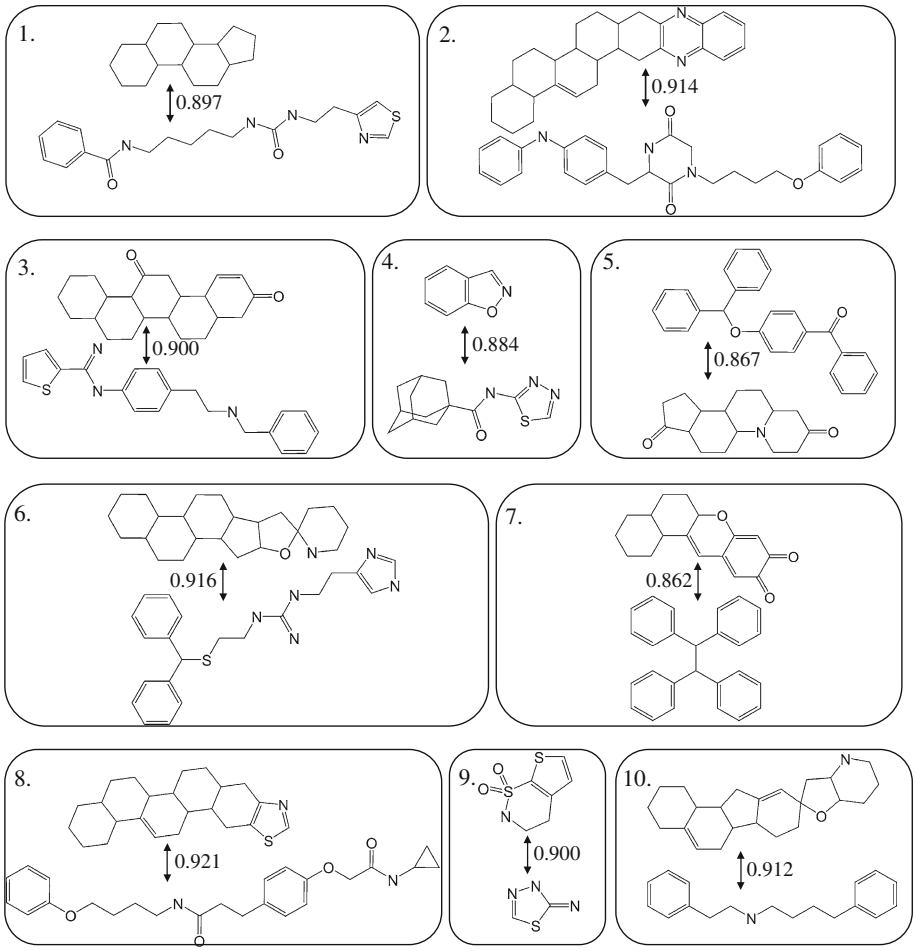

Fig. 4 Exemplary scaffold pairs. From each of the top 10 compound classes according to Table 1, an exemplary scaffold pair is shown that represents a challenge for scaffold hopping calculations. All scaffold pairs have a distance >0.850 and differ by no more than 10 non-hydrogen atoms in size

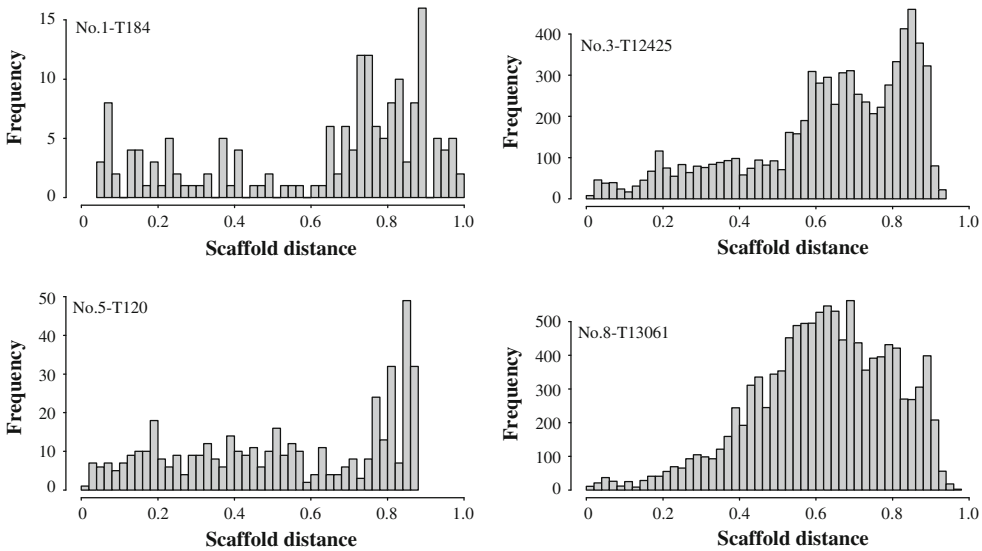

Fig. 5 Exemplary scaffold distance distributions. For four highly-ranked activity classes, intra-class scaffold distributions are reported. Ranks (No.) and TIDs are reported according to Table 1

order of 0.2, hence indicating a substantial spread of scaffold distances within activity classes, consistent with the presence of LD pairs and other more similar pairs with shorter distances.

Prioritization of activity classes

The top 10 classes in Table 1 contain between 23.4 and 8.0 % LD pairs and exemplary LD pairs from these

Table 2 Intra-class structural diversity

No	TID	No Cpds	No BMS	Cpds/BMS	No CSK	Cpds/CSK
1	184	80	19	4.2	12	6.7
2	10258	58	30	1.9	25	2.3
3	12425	252	119	2.1	68	3.7
4	12923	26	21	1.2	18	1.4
5	120	64	31	2.1	17	3.7
6	11165	36	31	1.2	25	1.4
7	12054	37	24	1.5	17	2.1
8	13061	336	153	2.2	122	2.7
9	10186	90	49	1.8	33	2.7
10	11574	74	26	2.8	20	3.7
11	12247	151	74	2.0	49	3.1
12	11398	57	41	1.4	31	1.8
13	11902	62	28	2.2	22	2.8
14	10553	124	59	2.1	26	4.7
15	12896	100	51	2.0	35	2.8
16	11060	154	67	2.3	40	3.8
17	11016	126	58	2.2	38	3.3
18	11042	144	65	2.2	43	3.3
19	20127	73	33	2.2	31	2.3
20	11232	59	37	1.6	30	1.9

For the top 20 activity classes according to Table 1, the ChEMBL target ID (TID), number of compounds per class (No Cpds), number of BM scaffolds (No BMS), the compound-to-scaffold ratio (Cpds/BMS), the number of cyclic skeletons (No CSK), and the compound-to-cyclic skeleton ratio (Cpds/CSK) are reported

classes are shown in Fig. 4. As can be seen, at large scaffold distances exceeding 0.850, scaffolds in LD pairs have only remote or partial structural resemblance or are virtually unrelated. Because LD pairs representing these largest structural distances comprise a maximum of ~20 % of all scaffold pairs, the activity classes contain many pairs of structurally more similar scaffolds (as indicated by average distances and standard deviations). This point is well illustrated in Fig. 5 that shows representative intra-class scaffold distance distributions for activity classes at ranks 1, 3, 5, and 8 in Table 1. As can be seen, the classes contain scaffold pairs yielding a continuum of small to large distances, with an enrichment of pairs yielding large distances. Thus, these activity classes represent attractive test cases for scaffold hopping calculations. In this context, it is also worth considering the number of scaffolds per class that are involved in the formation of LD pairs, as also reported in Table 1. For example, the top-ranked class, glutamate carboxypeptidase II inhibitors contained 19 unique scaffolds, giving rise to 171 possible pairs. All 19 scaffolds were involved in the formation of 40 LD pairs. Hence, 131 scaffold pairs had distances smaller than the LD threshold, yielding an average distance of 0.627. Furthermore, the class at rank 3, nitric oxide synthase inhibitors was larger and contained 119 scaffolds, 109 of which contributed to the formation of a total of 990 LD pairs. In addition to scaffold numbers, the size of ranked classes should also be considered. The four examples in Fig. 5 include two relatively small classes with fewer than 100 compounds (rank 1 and 5) and two larger ones with 252 (3) and 336 (8) compounds, all of which are of suitable size for LBVS test calculations. Because benchmarking of LBVS methods typically requires separating activity classes into subsets of reference molecules and active database compounds, the utility of small compound classes is often limited. However, only three of the 50 classes in Table 1 consisted of fewer than 50 compounds, while 30 classes contained more than 100 compounds (and three very large classes more than 1,000). Based on the ranking in Table 1 and the compound and scaffold information that is provided, activity classes can be prioritized and selected for specific investigations, considering relative LD pair content and additional criteria, as discussed above.

Structural diversity and similarity searching

For the top 20 activity classes in Table 1, molecular scaffold-based statistics are reported in Table 2 that mirror significant degrees of intra-class structural diversity, as to be expected. All classes are characterized by generally low compound-to-scaffold, scaffold-to-cyclic skeleton, and compound-to-cyclic skeleton ratios. Cyclic skeletons further abstract from scaffolds by converting all heteroatoms to carbon and setting all bond orders to one. Hence, cyclic skeletons represent topologically distinct scaffolds. In addition, Fig. 6 reports MACCS Tanimoto coefficient value distributions for these classes, which further mirror

Table 3 Similarity search performance

No	TID	MACCS				ECFP4			
		1-NN		10-NN		1-NN		10-NN	
		RR%	HR%	RR%	HR%	RR%	HR%	RR%	HR%
1	184	71.23	12.47	52.16	9.13	84.63	14.81	79.36	13.89
2	10258	41.45	4.97	27.85	3.34	81.35	9.76	71.73	8.61
3	12425	30.92	18.79	12.41	7.54	48.35	29.37	36.86	22.39
4	12923	28.56	1.14	8.06	0.32	29.78	1.19	26.31	1.05
5	120	68.69	9.27	47.91	6.47	74.73	10.09	61.65	8.32
6	11165	18.26	1.19	5.62	0.36	16.32	1.06	9.50	0.62
7	12054	40.27	2.82	31.04	2.17	38.93	2.73	34.43	2.41
8	13061	27.82	22.67	13.39	10.91	48.66	39.65	34.59	28.19
9	10186	16.81	3.36	5.41	1.08	26.10	5.22	2.80	4.16
10	11574	76.15	12.18	57.94	9.27	90.12	14.42	74.44	11.91
11	12247	44.24	16.81	9.67	3.67	73.35	27.87	58.14	22.09
12	11398	54.47	6.67	28.51	3.49	68.8	8.43	51.82	6.35
13	11902	58.70	7.63	25.58	3.33	71.92	9.35	59.04	7.67
14	10553	90.01	25.65	58.97	16.81	97.23	27.71	88.09	25.11
15	12896	21.97	4.94	5.14	1.16	28.76	6.47	2.60	4.63
16	11060	21.45	7.72	10.33	3.72	33.17	11.94	31.92	11.49
17	11016	21.94	6.36	10.86	3.15	31.45	9.12	29.58	8.58
18	11042	19.03	6.37	7.32	2.45	26.15	8.76	24.52	8.21
19	20127	42.48	6.69	17.27	2.72	68.69	10.82	58.10	9.15
20	11232	38.26	8.13	18.07	3.84	43.18	9.18	43.07	9.15

For the top 20 activity classes, results of fingerprint similarity search calculations are reported. For each class, recovery (RR%) and hit rates (HR%) were averaged over 100 independent trials

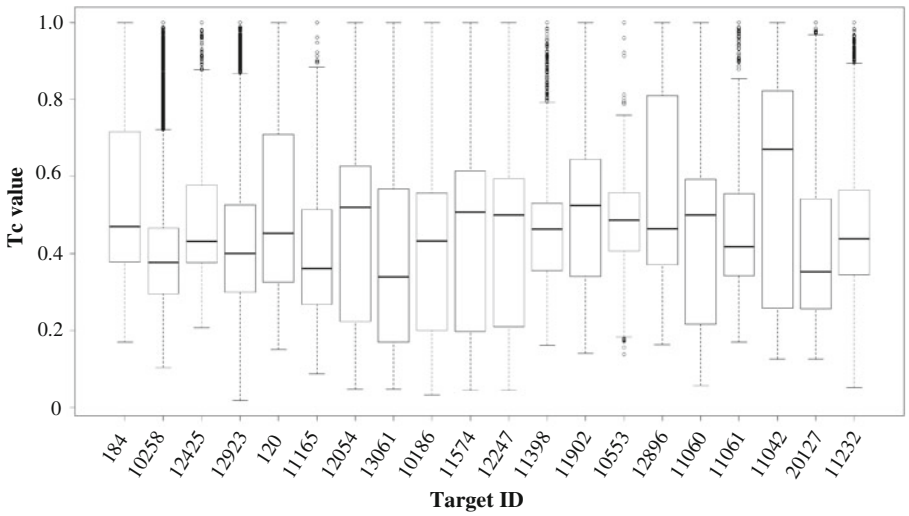

Fig. 6 Tanimoto similarity. For the top 20 activity classes according to Table 2, intra-class pair-wise MACCS Tanimoto coefficient (Tc) value distributions are reported as *box plots*. *Bars* indicate median values and the *upper* and *lower box boundaries* represent the first and third quartile, respectively. *Dotted lines* delineate the value range and *points above* and *below line boundaries* indicate outliers

intra-class structural diversity in most cases. Low compound-to-scaffold ratios also indicate that these classes are meaningful test cases for scaffold hopping analysis.

In Table 3, results of fingerprint similarity search control calculations are reported for the top 20 activity classes. The observed hit and recovery rate distribution reflects different levels of search performance for different classes, as one would expect. The rates also confirm that these classes are feasible for virtual screening applications. However, the relatively low hit and recall rates observed in many instances also indicate that these classes often present challenging test cases for scaffold hopping assessment,

at least for relatively simple similarity search approaches. These findings are in accord with the major goals of our analysis.

Conclusions

In this study, we have prioritized compound classes (for which high-confidence activity data was available) for the evaluation of the scaffold hopping potential of LBVS methods. It has been our intention to identify compound classes that represent meaningful and challenging test cases for scaffold hopping assessment. To these ends, inter-scaffold distances were systematically calculated for ChEMBL compound classes and a scaffold distance profile of the database was generated. On this basis, the scaffold information contained in all qualifying classes has been analyzed in detail. Activity classes were ranked according to the proportion of scaffold pairs with largest distances they contained. Up to $\sim$23 % of all scaffold pairs within a class fell into this category. The proportion of LD pairs was a meaningful indicator for the identification of activity classes with desirable scaffold distance distributions. Highly ranked classes had different sizes, mostly similar compound-to-scaffold ratios, contained different chemotypes, and covered different target families. The information provided herein should be helpful to select compound classes for scaffold hopping investigations. The 50 activity classes presented herein are freely available via the following URL: http://www.lifescienceinformatics.uni-bonn.de (see 'Downloads').

References

1. Schneider G, Neidhart W, Giller T, Schmid G (1999) Angew Chem Int Ed 19:2894
2. Schneider G, Schneider P, Renner S (2006) QSAR Comb Sci 25:1162
3. Brown N, Jacoby E (2006) Mini Rev Med Chem 6:1217
4. Hu Y, Bajorath J (2010) Med Chem Commun 1:339
5. Vogt M, Stumpfe D, Geppert H, Bajorath J (2010) J Med Chem 53:5707
6. Geppert H, Vogt M, Bajorath J (2010) J Chem Inf Model 50:205
7. Stumpfe D, Bajorath J (2011) In Sotriffer C (ed) Applied virtual screening: strategies, recommendations, and caveats. Methods and principles in medicinal chemistry. Virtual screening. Principles, challenges, and practical guidelines. Wiley-VCH, Weinheim, p 73
8. Hu Y, Stumpfe D, Bajorath J (2011) J Chem Inf Model 51:1742
9. Li R, Stumpfe D, Vogt M, Geppert H, Bajorath J (2011) J Chem Inf Model 51:2507
10. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP (2012) Nucleic Acids Res 40:D1100
11. ChEMBL (2012) http://www.ebi.ac.uk/chembldb/
12. Bemis GW, Murcko MA (1996) J Med Chem 39:2887
13. MACCS (2002) Structural keys. Symyx Software, San Ramon
14. Rogers D, Hahn M (2010) J Chem Inf Model 51:742
15. Irwin JJ, Shoichet BK (2005) J Chem Inf Model 51:177