J Comput Aided Mol Des (2013) 27:115–124
DOI 10.1007/s10822-012-9632-4

Activity cliffs in PubChem confirmatory bioassays taking inactive compounds into account

Ye Hu · Gerald M. Maggiora · Jürgen Bajorath

Received: 19 November 2012 / Accepted: 29 December 2012 / Published online: 8 January 2013

Abstract Activity cliffs are formed by pairs or groups of structurally similar compounds with significant differences in potency. They represent a prominent feature of activity landscapes of compound data sets and a primary source of structure–activity relationship (SAR) information. Thus far, activity cliffs have only been considered for active compounds, consistent with the principles of the activity landscape concept. However, from an SAR perspective, pairs formed by structurally similar active and inactive compounds should often also be informative. Therefore, we have extended the activity cliff concept to also take inactive compounds into consideration. As source of both confirmed active and inactive compounds, we have exclusively focused on PubChem confirmatory bioassays. Activity cliffs formed between pairs of active compounds (homogeneous pairs) and pairs of active and inactive compounds (heterogeneous pairs) were systematically analyzed on a per-assay basis, hence ensuring the currently highest possible degree of experimental consistency in activity measurement. Only very small numbers of large-magnitude activity cliffs formed between active compounds were detected in PubChem bioassays. However, when taking confirmed inactive compounds from confirmatory assays into account, the activity cliff frequency in assay data significantly increased, involving 11–15 % of all qualifying pairs of similar compounds, depending on the molecular representations that were used. Hence, these non-conventional activity cliffs provide an additional source of SAR information.

Keywords Active compounds · Compound potency · Inactive compounds · Molecular similarity · Activity cliffs

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

G. M. Maggiora
College of Pharmacy and BIO5 Institute, University of Arizona, 1295 North Martin, PO Box 210202, Tucson, AZ 85721, USA

G. M. Maggiora
Translational Genomics Research Institute, 445 North Fifth Street, Phoenix, AZ 85004, USA

Introduction

Activity cliffs are generally defined as pairs of structurally similar active compounds with a significant difference in potency [1, 2]. Given their "similar structure-different activity" characteristic, activity cliffs are of prime interest for structure–activity relationship (SAR) analysis, because they often reveal small structural modifications of active compounds that are SAR determinants. The description of activity cliffs depends on two critical parameters. First, it must be decided whether a continuum of activity cliffs with varying potency differences is considered in compound data sets [2, 3] or if a minimum potency difference between cliff partners is required [2, 4]. The latter approach focuses on cliffs of significant magnitude, but requires a clear- and consistently applied-definition of potency thresholds. Second, equally, if not more important, similarity of active compounds must be assessed- and quantified- as a prerequisite for activity cliff formation. The assessment of molecular similarity strongly depends on the chosen molecular representations (descriptors) and, to a lesser

extent, similarity measures [2, 5]. This introduces another variable in activity cliff assessment and, again, requires clear definition of similarity criteria. Accordingly, attempts have also been made to identify consensus activity cliffs that are consistently formed using different molecular representations (corresponding to different chemical reference spaces) [6].

The activity cliff concept has recently been extended in different ways. For example, given the strong molecular representation dependence of cliff formation, structure-based similarity criteria for activity cliffs have been introduced on the basis of the matched molecular pair (MMP) formalism [7], leading to the introduction of MMP-cliffs [8]. An MMP is defined as a pair of compounds that only differ by the exchange of a similar substructure at a single site, and MMPs with structural differences of limited size can be used to define activity cliffs (i.e., MMP-cliffs), thereby circumventing potential ambiguities associated with the use of alternative descriptors for similarity calculations. In addition, a structural categorization of activity cliffs has also been introduced, differentiating between five types of activity cliffs [9]. For example, according to this scheme, activity cliffs that only result from replacement of R-groups are distinguished from cliffs that involve core structure modifications or that are due to changes in stereochemistry [9]. Furthermore, activity cliff definitions on the basis of 3D structures of ligand–protein complexes have been introduced [10–12] and a survey of activity cliffs in the Protein Data Bank [13] has been reported. Finally, it has been observed that activity cliffs in compound data sets are often not formed in isolation (i.e., by pairs of compounds without structural neighbors), but rather by groups of structurally related compounds, leading to the notion of coordinated activity cliffs and the introduction of activity ridges [14, 15].

Although the activity concept has been extended in different ways, all current definitions and descriptions have in common that only pairs of active compounds are considered for the formation of cliffs, rather than active and inactive compounds. There are good reasons for this conventional view of activity cliffs, considering that they are central to the activity landscape concept [5]. An activity landscape is generally defined as any graphical representation that integrates the assessment of structure and potency relationships between active compounds [5], and activity cliffs are generally considered the most prominent and informative features of activity landscape models [5]. In principle, the inclusion of inactive compounds is problematic for activity landscape modeling because inactive compounds increase the size and the complexity of activity landscapes without necessarily adding information, which especially applies to structurally similar inactive compounds or pairs of inactive and weakly active ones. Hence, inactive compounds are usually not considered in activity landscape analysis. However, from an SAR perspective, also considering inactive compounds in activity cliff formation might provide additional information. Hence, if we focus on the analysis of activity cliffs outside of the conceptual framework of activity landscape representation, the inclusion of inactive compounds might be rather meaningful, although this has not been attempted thus far. Of course, doing so would require the availability of information about confirmed inactive molecules for a given target, which is not provided in major public domain medicinal chemistry databases such as BindingDB [16] or ChEMBL [17]. Typically, such information would be obtained from biological screening and confirmatory assay data, albeit associated with likely experimental variance.

Recently, we have carried out a survey of activity cliffs in the ChEMBL database [18] by exclusively considering active compounds using different molecular representations and similarity measures. As an activity cliff criterion, a potency difference between cliff partners of at least two orders of magnitude was required. The frequency of activity cliff formation across 414 target-specific compound data sets was determined. It was found that, on average, approximately 30 % of biologically active compounds were involved in the formation of activity cliffs. Herein, we report a proof-of-concept investigation to further extend activity cliff analysis by taking inactive compounds into account. Given the requirements discussed above and primary goals of our analysis, a systematic survey of confirmatory bioassays available in PubChem [19] was carried out, which yield both confirmed active compounds (screening hits) and confirmed inactive compounds in a given assay. Hence, the formation of activity cliffs was carried out on a per-assay basis for a total of 340 confirmatory bioassays, as reported herein. The results were also compared with those obtained on the basis of compound optimization data available in ChEMBL. The consideration of inactive compounds increases the frequency of activity cliffs and hence the source of SAR information. However, when only active PubChem compounds were considered, activity cliffs were much less frequent in screening data than in sets of optimized active compounds.

Materials and methods

Data collection

From the PubChem BioAssay database [19], confirmatory screening assays with defined activity measurements and dose–response data against a single protein target were

selected. For individual assays, all compounds designated as "active" or "inactive" were collected. Compounds denoted as "inconclusive" or "unspecified" were omitted. Compounds designated as active were only selected if explicit IC_{50} values were available. Our choice of confirmatory bioassay data is further rationalized below.

Similarity evaluation

For the assessment of 2D structural similarity, three different molecular representations were calculated including MACCS structural keys [20], the extended connectivity fingerprint with bond diameter of four (ECFP4) [21], and matched molecular pairs (MMPs) [7]. MACCS and ECFP4 represent two types of 2D fingerprints of different designs. The former consists of a total of 166 predefined structural fragments, whereas the latter is a topological fingerprint encoding layered atom environments up to a diameter of four bonds around each atom in a compound. The MACCS and ECFP4 similarity of compounds was quantified using the Tanimoto coefficient (Tc) [22].

MMPs were generated using an in-house implementation of the algorithm by Hussain and Rea [7]. Compounds forming an MMP only differ at a single site by the exchange of a pair of substructures of only limited size, referred to a transformation [7]. Following transformation size criteria introduced for the definition of MMP-cliffs [8], MMP formation was restricted as follows:

1. For each compound, the size of the exchanged substructure was permitted to be at most half the size of the remaining core fragment.
2. The difference in the size between the two exchanged substructures was limited to maximally eight non-hydrogen atoms.
3. The maximal size of an exchanged substructure was set to 13 non-hydrogen atoms.

Activity cliff criteria

The following criteria were applied for the formation of activity cliffs:

1. For MACCS and ECFP4, the similarity thresholds were set to 0.85 and 0.55, respectively. For the two fingerprints, these Tc values correspond to each other [23]. In addition, compounds were considered to meet the activity cliff similarity criterion if they formed a transformation size-restricted MMP [8].
2. For pairs of similar active compounds that were considered homogeneous pairs, an at least 100-fold difference in potency (IC_{50} values) was required as an activity cliff criterion.
3. All similar compound pairs formed between an active and an inactive compound, considered heterogeneous pairs, were in the first instance counted as activity cliffs. To further refine the selection and ensure that very weakly active compounds were not considered in combination with inactive ones, the subset of active/inactive cliffs was determined in which the active compound had at least 10 μM potency.

Pairs of inactive compounds were not considered in our analysis. Activity cliff distributions were compared with those recently obtained for ChEMBL [18].

Results and discussion

Focusing on confirmatory bioassays

In our analysis, we have deliberately focused on confirmatory assay data, for several reasons. First and foremost, activity cliff analysis heavily relies on accurate compound activity data, which generally limits the use of original screening data. Hence, the availability of confirmed active compounds with accurate activity measurements is essential. Moreover, in order to include information provided by inactive compounds into activity cliff analysis, the availability of confirmed inactives is also a must. Hence, given stringent data integrity requirements, especially for large-scale analysis, confirmatory assays represented the premier source of compound activity data from screening experiments, although they provided an incomplete account of inactive compounds in a given assay (the majority of which are not considered for confirmatory experiments). In order to take all inactive compounds into consideration, not further evaluated inactives from a primary screen might also be added to confirmatory assay sets. However, for the purpose of our investigation, this was not required and also prohibitive, for the following reasons. Clearly, for reliable activity cliff assignments, both false-positives and false-negatives must be eliminated (to the extent possible). Importantly, compound selection from primary screening data for confirmatory assays in PubChem is not consistent and represents a major additional variable for large-scale analysis (see the example given below). Moreover, as reported in Table 1, confirmatory bioassays selected under stringent criteria were found to contain overall more confirmed inactive than active compounds (10,103 vs. 8,255, respectively). As further shown below, these compounds were structurally diverse. Thus, the data set was well balanced and provided a sound basis for investigating the influence of inactive compounds on the rate of activity cliff formation, at the highest possible level of assay data confidence, while avoiding a very strong additional bias towards inactive compounds. As a

Table 1 Data composition

No. of	PubChem bioassays	ChEMBL
Compounds		
Active	8,255	27,610
Inactive	10,103	–
Active/inactive	5,700	–
Total	24,058	27,610
Assays	340	3988
Targets	143	414
Type of activity measurement	IC_{50}	K_i
Compound-assay combinations	48,679	45,139

For PubChem bioassays (confirmatory), the number of qualifying active and/or inactive compounds, assays, and targets are reported. "Active/inactive" means that compounds are active or inactive in different assays. In addition, the number of unique combinations of compounds and assays is reported (e.g., if a given compound is tested in three assays, three unique compound-assay combinations are obtained). For comparison, the numbers of active compounds, targets, assays, and assay-target combinations are reported for ChEMBL (release 10) according to Ref. [18]. It should be noted that the ChEMBL activity cliff survey was carried out on the basis of individual target sets and (assay-independent) equilibrium constants (K_i values), while PubChem only reports IC_{50} values

representative example, the confirmatory assay for agonists of transient receptors potential channels 3 (TRPML3; PubChem assay ID 602129) contained 12 confirmed active and 22 confirmed inactive compounds, whereas the primary parent screen (assay ID 1448) reported 632 putatively active and 217,332 inactive compounds. Thus, including a large amounts of inactive compounds from such assays would have been prohibitive for our large-scale analysis. In addition, for practical applications, results provided on the basis of confirmatory assays are particularly relevant because activity cliffs are mostly explored in the context of SAR analysis where limited numbers of inactive (or nearly inactive) compounds are available. Hence, taken together, the exclusive use of confirmatory assay data provided the most reliable (and feasible) basis for our current analysis.

Assay data

On the basis of our selection criteria, a total of 1,088 confirmatory screening assays with defined activity measurements and dose–response behavior against individual target proteins were collected from the PubChem BioAssay database (July 2012 release). For 340 of these assays, which corresponded to 143 unique targets, active compounds were annotated with explicitly defined IC_{50} values and were selected for further analysis. From these 340 assays, both confirmed active and inactive compounds were retrieved, yielding a total of 24,058 compounds representing 48,679 unique compound-assay combinations, as reported in Table 1. Approx. 34 and ~42 % of these 24,058 compounds were exclusively designated as active or inactive compounds in one or more assays, whereas the remaining 24 % (i.e., 5,700 compounds) were active or inactive in different assays. Table 1 also reports the corresponding statistics for ChEMBL release 10 for which activity cliff distributions were previously determined [18]. The number of compounds and compound-assay combinations were comparable to those obtained for PubChem, due to the inclusion of confirmed inactive compounds. However, compounds and activity data in ChEMBL were associated with a much larger number of assays and targets.

Compound distribution

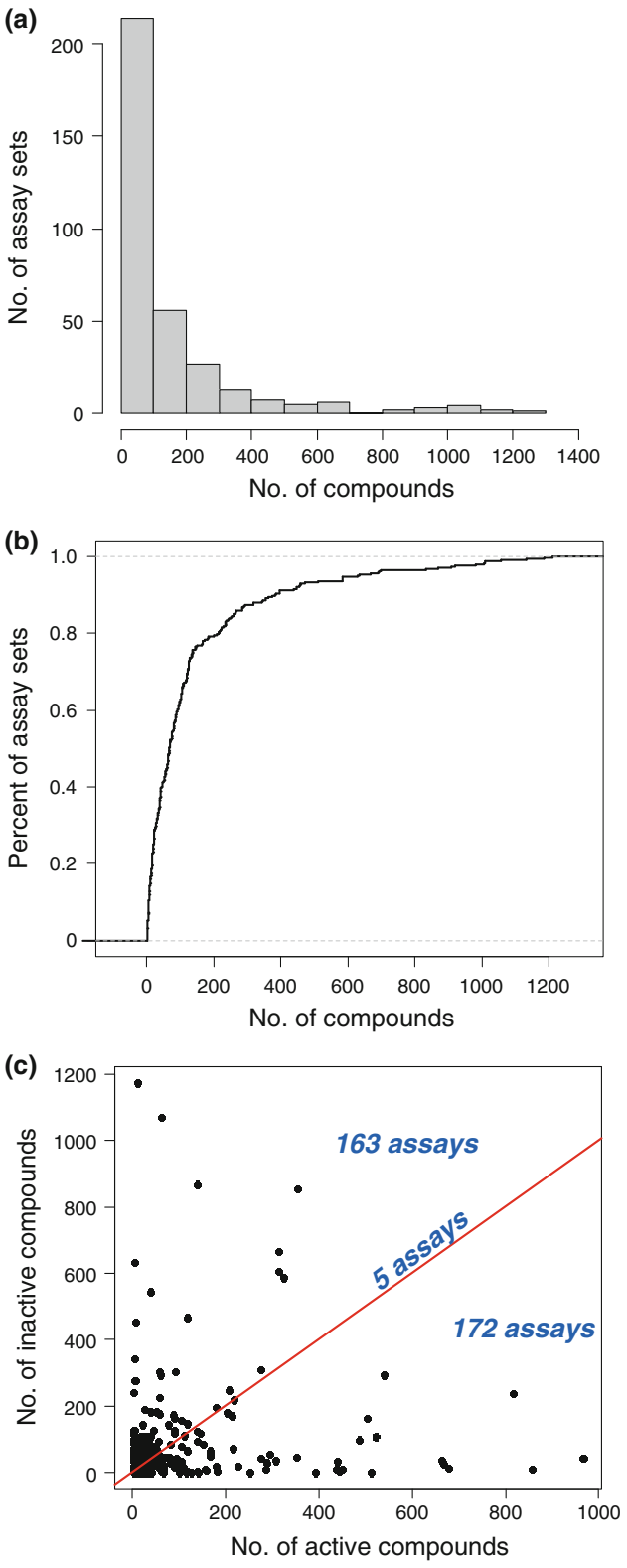
The 340 selected PubChem assays contained between one and 1,210 compounds. Their compound distribution is reported in Fig. 1a, b. The majority of assays (i.e., 212) consisted of fewer than 100 compounds. Furthermore, the distribution of active versus inactive compounds is reported in Fig. 1c. A total of 172 assays below the diagonal contained more active than inactive compounds. By contrast, 163 assays above the diagonal contained more inactive compounds. In addition, five assays had the same number of active and inactive compounds.

Potency value distribution

Previously, ChEMBL compounds were assigned to five different potency subranges spanning 12 orders of magnitude, from picomolar to molar potency [18]. This classification scheme was adopted to analyze the global potency value distribution of active PubChem compounds, as reported in Table 2. Nearly all potency values fell into the millimolar (>64 %) and micromolar (>35 %) subranges, as one would expect for screening data. Only 16 nM potency records (0.06 %) were detected. By contrast, more than 23 % of the potency values of ChEMBL compounds fell into the nanomolar subrange, consistent with their primary origin (i.e. compound optimization data).

Molecular similarity

For each PubChem assay, compounds were compared in a pairwise manner. The distribution of Tc values calculated using MACCS keys and ECFP4 is reported in Fig. 2a. Among all possible 11,477,970 compound pairs, the majority yielded MACCS Tc of ≤0.5 and ECFP4 Tc of ≤0.2. A total of 52,171 pairs (~0.45 %) and 49,868 pairs (~0.43 %) yielded MACCS Tc of at least 0.85 and ECFP4 Tc of at least of 0.55, respectively. In addition, 32,065 pairs (~0.28 %) formed MMPs. Furthermore, no qualifying compound pairs



◀ **Fig. 1** Compound distribution. **a** Shown is the distribution of the number of compounds for all 340 qualifying PubChem confirmatory assays. The corresponding cumulative distribution is shown in (**b**). In **c**, for each assay, the number of active and inactive compounds is compared in a *scatter plot* format. Each *dot* represents an assay. A total of 172 assays below the diagonal contained more active than inactive compounds, 163 assays above the diagonal contained more inactive compounds, and five assays had the same number of active and inactive compounds

Table 2 Potency value distribution

Potency value subranges		No. of potency values (%)	
		PubChem bioassays	ChEMBL
(11, 14]	>10–≤0.01 pM	0	31 (0.07)
(8, 11]	>10–≤0.01 nM	16 (0.06)	10.560 (23.39)
(5, 8]	>10–≤0.01 μM	16,244 (64.67)	31,047 (68.78)
(2, 5]	>10–≤0.01 mM	8,859 (35.27)	3,370 (7.47)
(-1, 2]	>10–≤0.01 M	0	131 (0.29)
Total		25,119	45,139

For PubChem assays and ChEMBL data sets, the number of potency values falling into each of five subranges (from picomolar to molar potency) and the corresponding ratio (in percent) are reported

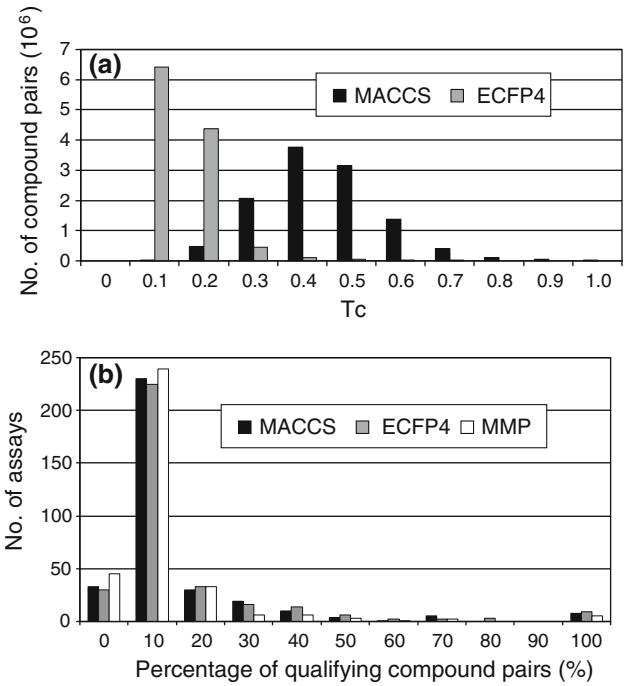


Fig. 2 Compound similarity. **a** The distribution of Tanimoto coefficient (Tc) values is reported for all possible pairs of assay compounds calculated using MACCS structural keys (*black*) and the ECFP4 fingerprint (*gray*), respectively. **b** For all qualifying PubChem confirmatory assays, the ratio of the compound pairs that yielded MACCS Tc values of at least 0.85 (*black*), ECFP4 Tc values of at least of 0.55 (*gray*), or MMPs (*white*) is reported

were observed in 30–45 assay sets, depending on the molecular representation, as shown in Fig. 2b. However, irrespective of the molecular representations, less than 10 % of the compound pairs were considered as structurally similar for majority of the assay sets. Thus, assay/screening set compounds were structurally diverse in many instances.

Table 3 Classification of compounds pairs and activity cliffs

Categorization	No. of					
	MACCS Tc ≥ 0.85		ECFP4 Tc ≥ 0.55		MMP	
	Pairs	Cliffs	Pairs	Cliffs	Pairs	Cliffs
Active/active (homogeneous)						
mM/mM	2,948	0	2,874	0	1,931	0
μM/mM	4,922	45	4,381	36	2,904	22
μM/μM	14,316	3	14,299	5	9,845	10
μM/nM	34	16	43	19	16	5
nM/nM	5	0	5	0	4	0
Active/inactive (heterogeneous)						
mM/inactive	4,494	4,494	3,655	3,655	1,961	1,961
μM/inactive	7,840	7,840	6,627	6,627	3,485	3,485
nM/inactive	4	4	8	8	2	2
Inactive/inactive	17,608	–	17,976	–	11,917	–
Total	52,171	**12,402**	49,868	**10,350**	32,065	**5,485**

Different compound pair categories were defined taking different potency subrange combinations and inactive compounds into account. For each category, the number of compound pairs exceeding similarity threshold values for MACCS and ECFP4 and the number of compound pairs forming MMPs are given. For eight pair categories, the corresponding numbers of activity cliffs are reported. By definition, "Inactive/inactive" pairs cannot form activity cliffs

Activity cliff frequency

Given our criteria for the formation of activity cliffs, taking both active and inactive compounds into account, a total of 12,402, 10,350, and 5,485 activity cliffs were identified for the MACCS, ECFP4 and MMP representations, respectively (Table 3). These cliffs corresponded to ~23.77, ~20.75, and 17.11 % of all qualifying pairs of similar compounds for these representations. Therefore, less than one-fourth of structurally similar compound pairs formed activity cliffs and more fingerprint-based activity cliffs were formed than MMP-cliffs, as also observed for ChEMBL compounds [18].

Activity cliff classification

All qualifying compound pairs and activity cliffs were assigned to eight different categories on the basis of activity designations and potency subranges into account, as reported in Table 3. For pairs of similar active compounds (i.e., homogeneous pairs), five categories were defined that accounted for different potency subrange combinations. For qualifying pairs formed between an active and inactive compound (i.e., heterogeneous pairs), three categories were defined by considering the potency subrange of the active compound. In addition, pairs of similar inactive compounds were also counted.

For pairs exceeding the MACCS Tc threshold of 0.85, ~42.60 and ~23.65 % were formed by homogeneous and heterogeneous pairs, respectively. However, only 64 of 22,225 homogeneous pairs displayed an at least 100-fold difference in potency and hence qualified as activity cliffs according to our criteria. These activity cliffs fell into three categories, i.e., 45 cliffs were formed by compounds with μM/mM potency, three cliffs by μM/μM compounds, and 16 cliffs by μM/nM compounds. However, when all qualifying heterogeneous pairs were considered, 12,338 additional cliffs were formed for the MACCS representation, often involving weakly active (μM) compounds. Comparable observations were made ECFP4 and MMP-cliffs. Only 60 and 37 cliffs were formed by homogeneous

Table 4 Activity cliff statistics

	No. of (%)					
	PubChem bioassays			ChEMBL		
	MACCS	ECFP4	MMP	MACCS	ECFP4	MMP
Qualifying pairs	52,171	49,868	32,065	314,665	316,034	198,569
Activity cliffs						
Active/active (ΔpIC50 ≥ 2)	64 (0.12)	60 (0.12)	37 (0.12)	17,886 (5.68)	16,941 (5.36)	9,481 (4.77)
Active/active (ΔpIC50 ≥ 2) and active (≥10 μM)/inactive	7,923 (15.19)	6,704 (13.44)	3,530 (11.01)	–	–	–
Total	12,402 (23.77)	10,350 (20.75)	5,485 (17.11)	17,886 (5.68)	16,941 (5.36)	9,481 (4.77)
Cliff-forming compounds	3,984 (16.56)	4,145 (17.23)	3,483 (14.25)	9,520 (34.48)	9,159 (31.70)	6,286 (22.77)

For PubChem assays and ChEMBL compound activity classes, the numbers of qualifying compound pairs (exceeding Tanimoto similarity thresholds or meeting the MMP similarity criterion) are reported. In addition, the corresponding number (ratio) of cliffs formed between active compounds with at least 100-fold difference in potency is given. Furthermore, for PubChem assays, activity cliffs were determined that were formed by an inactive compound and an active compound with at least 10 μM potency. Pairs of inactive and very weakly active compounds were not considered here as potential cliffs. In addition, the number (ratio) of compounds involved in the formation of all activity cliffs (Cliff-forming compounds) is given

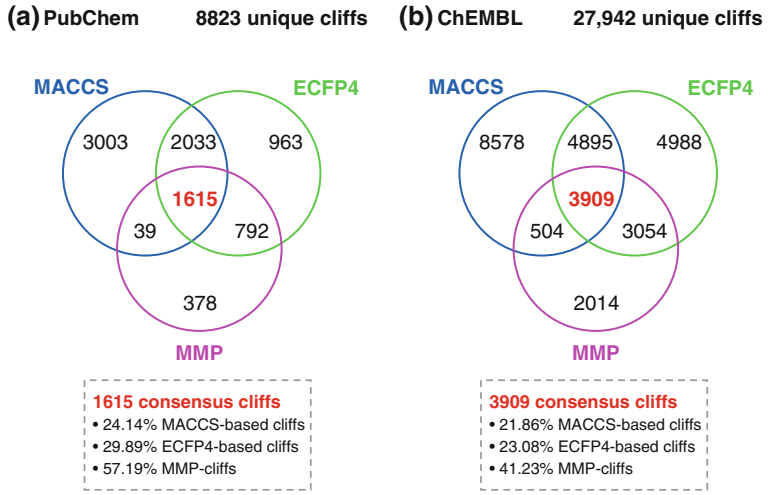


Fig. 3 Activity cliff overlap. Activity cliffs formed by **a** PubChem and **b** ChEMBL compounds are compared in *Venn diagram* representations. Activity cliffs were defined on the basis of MACCS or ECFP4 Tanimoto similarity or on the basis of the MMP formulism. For ChEMBL activity cliffs, a 100-fold potency difference between cliff partners was required. For PubChem activity cliffs formed between active compounds, the same potency difference criterion was applied. In addition, for PubChem cliffs formed between an active and inactive compound, the active molecule was required to have a potency value of at least 10 μM. For each type of activity cliff, the proportion of consensus cliffs is reported

compound pairs for ECFP4 and MMPs, respectively, but more than 10,000 additional cliffs by heterogeneous pairs for ECFP4 and more than 5,000 additional MMP-cliffs.

Activity cliff refinement

To further refine the activity cliff definition and ensure that very weakly active compounds were not considered in combination with inactive ones, cliffs formed by heterogeneous compound pairs were further restricted to those for which the active compounds had a potency of at least 10 μM. As reported in Table 4, the total number of activity cliffs was then reduced to 7,923, 6,704, and 3,530 for MACCS, ECFP4 and MMP representations, respectively. These cliffs corresponded to ∼15.19, ∼13.44 and ∼11.01 % of all qualifying compound pairs.

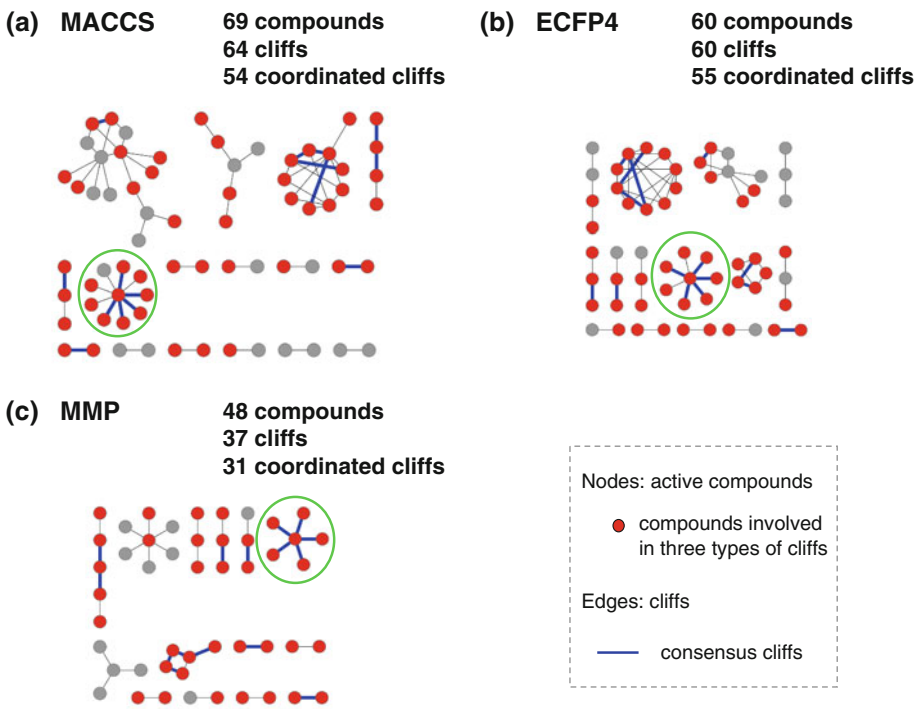


Fig. 4 Activity cliff networks. Activity cliffs formed by active compounds having potency difference of at least two orders of magnitude are visualized in networks for different molecular representations including **a** MACCS, **b** ECFP4, and **c** MMP. In each network, *nodes* represent compounds that are connected by an *edge* if they form an activity cliff. Compounds involved in the formation of all three types of cliffs (but not necessarily the same cliffs) are colored *red*. In addition, consensus cliffs consistently formed for all three molecular representations are indicated by *bold blue edges*. The cluster containing the largest number of consensus cliffs is encircled (*green*)

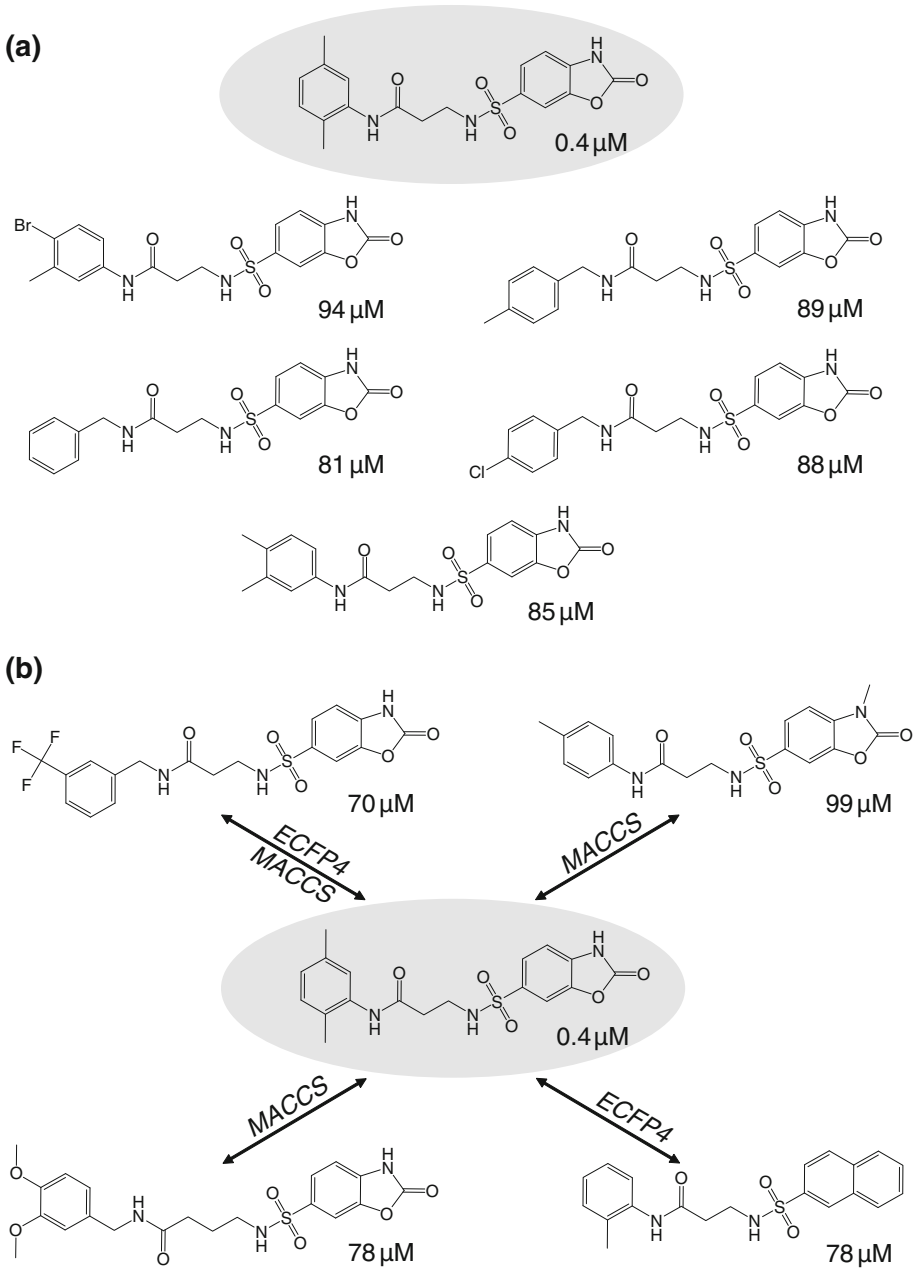


Fig. 5 Representative activity cliffs. In the encircled consensus cliff clusters in Fig. 4, five consensus cliffs were formed involving the same compound (shown on a *gray* background). In **a**, the five cliff partners are shown. The IC_{50} values of these compounds for the mouse intestinal alkaline phosphatase are reported. In **b**, four other (non-consensus) activity cliffs from the same assay are shown that also involved the same central compound

Activity cliff overlap

These restricted activity cliff populations for different molecular representations were compared, as reported in Fig. 3a. A total of 3,003, 963, and 378 activity cliffs were exclusively formed by MACCS, ECFP4, and MMPs, respectively. However, a total of 1,615 activity cliffs were consistently detected for all three molecular representations and hence considered consensus cliffs. These cliffs represented ~24.14, ~29.89, and ~57.19 % of all MACCS-, ECFP4-, and MMP-cliffs, respectively. In total, 8,823 unique activity cliffs were observed. For comparison, Fig. 3b reports the activity cliff overlap for ChEMBL compounds. In this case, larger numbers of activity cliffs and consensus cliffs were detected. However, the proportion of consensus cliffs was comparable to PubChem.

Activity cliff networks

For the 64, 60, and 37 activity cliffs exclusively formed by homogeneous pairs for MACCS, ECFP4, and MMP, respectively, network representations were designed to visualize relationships between cliff forming compounds, as shown in Fig. 4. In each network, nodes represented active compounds and edges indicated activity cliffs. Compounds involved in the formation of all three types of

cliffs (but not necessarily the same cliffs) were colored red. In addition, consensus cliffs were depicted using bold blue edges. Although the topology of networks varied for different representations, small clusters of activity cliffs were consistently observed. Five consensus cliffs formed a cluster in all cases for the set of phosphatase inhibitors. These five consensus cliffs involved the same central compound and are shown in Fig. 5a. The structural differences between the central compound and its cliff partners originated from substitutions of the benzene ring and the position of the amide group, which led to potency alterations of more than two orders of magnitude. In Fig. 5b, four non-consensus cliffs from the same assay set are shown that involved the same central compound. These cliffs were only identified for MACCS and/or ECFP4.

Conclusions

Herein, we have presented an extension of the activity cliff concept by taking inactive compounds into account. Given the chemical diversity of many compound sets used for biological screening, including information provided by inactive compounds into activity cliff and SAR analysis is of high practical relevance. As a source of confirmed active and inactive compounds, PubChem confirmatory bioassays were thoroughly analyzed on a per-assay basis. The exclusive focus on confirmatory assay data was in line with the major goals of our analysis (although confirmatory assay data provide an incomplete account of inactives in a given assay). Large-magnitude activity cliffs between active compounds were only rarely formed in these assays, consistent with their screening data nature and relatively narrow potency distributions, with a dominance of weakly active compounds. Hence, there was only very little SAR information available in the form of large-magnitude activity cliffs. However, when inactive compounds were taken into account, large numbers of non-conventional activity cliffs were observed. We restricted the formation of non-conventional cliffs formed between active/inactive compounds to those for which the active compound had a potency of at least 10 μM, which still yielded between ~3,500 and nearly 8,000 activity cliffs for 340 qualifying assays, depending on the molecule representation that was used. MMP-cliffs provided a structurally much more conservative account of activity cliffs than calculated fingerprint Tanimoto similarities, in accord with previous observations. Hence, by taking into consideration a limited number of confirmed inactive compounds, the activity cliff knowledge base was substantially increased. Although large numbers of non-conventional activity cliffs identified in PubChem might not be suitable for activity landscape modeling, they do provide an additional source of SAR information, in particular, for screening data where starting points for SAR exploration are often difficult to identify. As such, information provided by activity cliffs formed between active/inactive compounds should further complement current SAR analysis approaches.

References

1. Maggiora GM (2006) On outliers and activity cliffs—why QSAR often disappoints. J Chem Inf Model 46(4):1535
2. Stumpfe D, Bajorath J (2012) Exploring activity cliffs in medicinal chemistry. J Med Chem 55(7):2932–2942
3. Guha R, Van Drie JH (2008) Structure-activity landscape index: identifying and quantifying activity cliffs. J Chem Inf Model 48(3):646–658
4. Wassermann AM, Dimova D, Bajorath J (2011) Comprehensive analysis of single- and multi-target activity cliffs formed by currently available bioactive compounds. Chem Biol Drug Des 78(2):224–228
5. Wassermann AM, Wawer M, Bajorath J (2010) Activity landscape representations for structure—activity relationship analysis. J Med Chem 53(23):8209–8223
6. Medina-Franco JL, Martínez-Mayorga K, Bender A, Marín RM, Giulianotti MA, Pinilla C, Houghten RA (2009) Characterization of activity landscapes using 2D and 3D similarity methods: consensus activity cliffs. J Chem Inf Model 49(2):477–491
7. Hussain J, Rea C (2010) Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. J Chem Inf Model 50(3):339–348
8. Hu X, Hu Y, Vogt M, Stumpfe D, Bajorath J (2012) MMP-cliffs: systematic identification of activity cliffs on the basis of matched molecular pairs. J Chem Inf Model 52(5):1138–1145
9. Hu Y, Bajorath J (2012) Extending the activity cliff concept: structural categorization of activity cliffs and systematic identification of different types of cliffs in the ChEMBL database. J Chem Inf Model 52(7):1806–1811
10. Seebeck B, Wagener M, Rarey M (2011) From activity cliffs to target-specific scoring models and pharmacophore hypotheses. ChemMedChem 6(9):1630–1639
11. Hu Y, Bajorath J (2012) Exploration of 3D activity cliffs on the basis of compound binding modes and comparison of 2D and 3D cliffs. J Chem Inf Model 52(3):670–677
12. Hu Y, Furtmann N, Gütschow M, Bajorath J (2012) Systematic identification and classification of three-dimensional activity cliffs. J Chem Inf Model 52(6):1490–1498
13. Berman H, Henrick K, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28(1):235–242
14. Vogt M, Huang Y, Bajorath J (2011) From activity cliffs to activity ridges: informative data structures for SAR analysis. J Chem Inf Model 51(8):1848–1856
15. Namasivayam V, Bajorath J (2012) Searching for coordinated activity cliffs using particle swarm optimization. J Chem Inf Model 52(4):927–934
16. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res 35(Database issue):D198–D201
17. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP (2012) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40(Database issue):D1100–D1107

18. Stumpfe D, Bajorath J (2012) Frequency of occurrence and potency range distribution of activity cliffs in bioactive compounds. J Chem Inf Model 52(9):2348–2353
19. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Zhou Z, Han L, Karapetyan K, Dracheva S, Shoemaker BA, Bolton E, Gindulyte A, Bryant SH (2012) PubChem's bioassay database. Nucleic Acids Res 40(Database issue):D400–D412
20. MACCS Structural Keys, Symyx Software: San Ramon, CA, 2005
21. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50(5):742–754
22. Willett P (2005) Searching techniques for databases of two- and three-dimensional structures. J Med Chem 48(13):4183–4199
23. Wawer M, Bajorath J (2010) Similarity-potency trees: a method to search for SAR information in compound data sets and derive SAR rules. J Chem Inf Model 50(8):1395–1409