JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

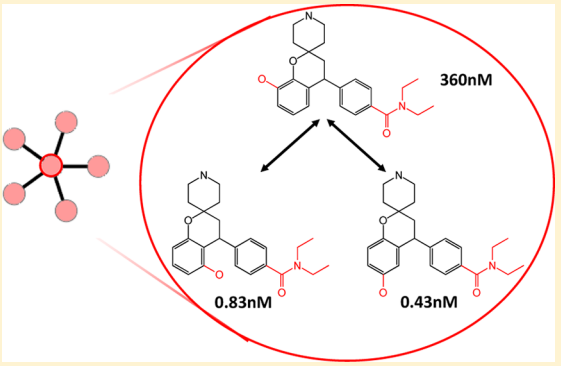
pubs.acs.org/jcim

Extending the Activity Cliff Concept: Structural Categorization of Activity Cliffs and Systematic Identification of Different Types of Cliffs in the ChEMBL Database

Ye Hu and Jürgen Bajorath*

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

ABSTRACT: Activity cliffs are generally understood as pairs or groups of similar compounds with large differences in potency. The study of activity cliffs is of high interest for the characterization of activity landscapes of compound data sets and the identification of SAR determinants, given the "small chemical change(s)−large potency effect" phenotype of cliffs. Herein, we introduce a new structural classification scheme for activity cliffs and introduce new cliff types. Activity cliffs are divided into five different classes dependent on whether the participating compounds are only distinguished by chirality, topology, R-group sets, core structures (scaffolds), or core structures and R-group topology. All cliff types are frequently detected in the ChEMBL database. R-group cliffs occur with higher propensity than other cliff types, as one might expect. However, many scaffold and R-group cliffs are not identified on the basis of whole-molecule similarity calculations, although they are often chemically intuitive. This makes the activity cliff classification attractive for medicinal chemistry analysis, independent of similarity calculations. Assignment of activity cliffs on the basis of well-defined structural criteria complements and further extends current approaches to identify and represent cliffs.



INTRODUCTION

The activity cliff concept experiences increasing interest in chemoinformatics and medicinal chemistry.[1−4] Supported by advances in activity landscape modeling and SAR visualization,[2−4] activity cliffs have increasingly become a focal point of SAR analysis. Moreover, they are also of interest in compound data mining.[4,5] In fact, through data mining, the frequency of activity cliff formation in bioactive compounds has been determined,[5] and it has been established that cliffs are often formed in a coordinated manner,[6,7] leading to the notion of activity ridges.[6] To define activity cliffs, potency difference thresholds and similarity criteria must be specified for cliff forming compounds.[4] These criteria are critical for activity cliff assessment, but their choice is often subjective. Hence, cliffs can be defined in rather different ways, for example, as discrete potency difference/similarity states[4,5] or as a continuum of states.[4,8] Moreover, there are many alternative ways to represent active compounds and assess their similarity. Regardless of how activity cliffs are defined, their distributions are also affected by variability and confidence levels of potency measurements[9] and, in addition, by the potency ranges that are considered for strongly active and weakly active cliff partners.[4,5] However, the way compound similarity is assessed presents the most critical variable for activity cliff analysis.[3,4] Here, chosen molecular representations and the way their similarity is quantified play a key role. Early investigations identified cliff partners on the basis of whole-molecule similarity calculations,[2−4] mostly by calculating Tanimoto similarity[10] for representations such as MACCS structural keys[11] or extended connectivity fingerprints.[12] In these cases, Tanimoto similarity threshold values were subjectively set, representing a major variable for cliff detection. In addition to this conventional assessment of similarity on the basis of fingerprints or other descriptors, three-dimensional activity cliffs have also been introduced by comparing the similarity of binding modes of ligands in complex X-ray structures.[13] Such comparisons require the application of three-dimensional similarity metrics.[13,14] Furthermore, as an alternative to calculated similarity values, substructure relationships have been used as a similarity criterion. This has been accomplished through the application of the matched molecular pair (MMP) formalism.[15] An MMP is defined as a pair of compounds that only differ at a single site. By only considering chemical differences of small size, activity cliffs have been defined and chemical transformations identified that display a tendency to form activity cliffs across different target families.[16,17]

Herein, we introduce a conceptually different approach to define activity cliffs that considers molecular scaffolds, R-group patterns, topology, and chirality as criteria for a systematic classification of activity cliffs. This classification scheme goes beyond substructure relationships, enables the assessment of

Received: June 15, 2012
Published: July 4, 2012

dx.doi.org/10.1021/ci300274c | J. Chem. Inf. Model. 2012, 52, 1806−1811

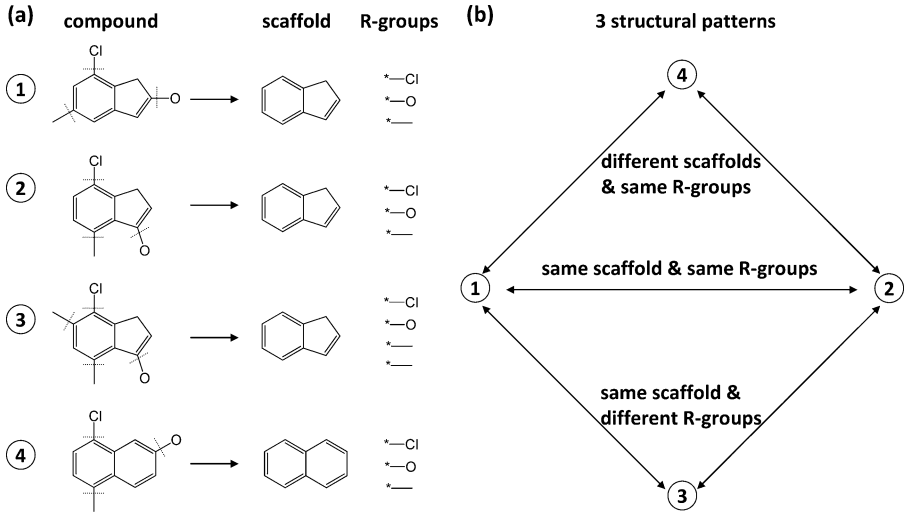


Figure 1. Compound decomposition. (a) Shown are four compounds decomposed into scaffolds and R-groups. (b) Different pairs of these four compounds are characterized by well-defined structural relationships. Compounds 1 and 2 contain indene as their scaffold and share three R-groups (category 1); compound 3 shares the same scaffold with compounds 1 and 2 and contains an overlapping, yet distinct set of R-groups (category 2); compound 4 has the same R-group set as compounds 1 and 2, but contains a different (naphthalene) scaffold (category 3).

activity cliff formation at different structural levels, and yields a further refined representation of activity cliffs that supports chemical interpretation.

MATERIALS AND METHODS

Compound Data. From ChEMBL (release 13),[18,19] compounds reported to directly interact (i.e., target relationship type "D") with human targets at the highest confidence level (i.e., target confidence score 9) with available K_i measurements were assembled. Compounds with multiple K_i values against the same target that differed by more than 1 order of magnitude were discarded to ensure a high level of data consistency. For multiple K_i measurements within the same order of magnitude, the geometric mean was calculated as the final potency annotation. All qualifying compounds were organized into individual target sets. Only target sets containing at least two compounds were retained for further analysis.

Compound Decomposition. Following Bemis and Murcko,[20] each target set compound was decomposed into two structural components, i.e., its molecular scaffold and R-groups, as illustrated in Figure 1a. On the basis of this decomposition scheme, structural differences between compounds can be assessed by separately comparing their scaffolds and R-groups. Because of its abundance in public domain compounds, the most generic scaffold, a single benzene ring, was omitted from further consideration. This avoided statistical bias in the evaluation of activity cliff distributions.

Structural Categorization. Compounds were initially assigned to pairs on the basis of shared scaffolds and/or R-groups. Accordingly, three categories indicating different degrees of structural relatedness were defined:

(1) Two compounds share the same scaffold and set of R-groups.
(2) Two compounds contain the same scaffold, but different sets of R-groups.
(3) Two compounds contain different scaffolds, but the same set of R-groups.

These categories are illustrated in Figure 1b. Different sets of R-groups might be completely distinct or partly overlapping.

For compound pairs falling into structural categories 2 and 3, additional selection criteria were applied, as illustrated in Figure 2. In accord with the activity cliff concept, in order to limit

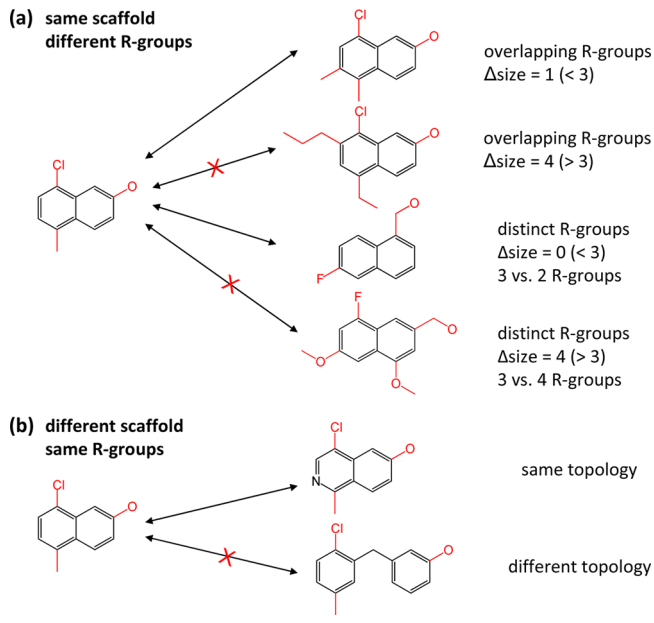


Figure 2. Compound pair selection criteria. (a) Compound pairs containing the same scaffold and different sets of R-groups were only retained if the total size difference between their R-group sets was smaller than three non-hydrogen atoms. In addition, each participating compound was permitted to maximally contain three different R-groups. (b) Compound pairs with different scaffolds and identical R-groups were only selected if scaffolds were topologically equivalent.

structural differences in compounds forming pairs in category 2, the size difference of R-group sets was limited to less than three non-hydrogen atoms. Moreover, each participating compound was permitted to contain a maximum of three different R-groups. In addition, compound pairs in category 3 had to contain topologically equivalent scaffolds, i.e., scaffolds yielding the same cyclic skeleton.[21] This ensured that scaffolds were

dx.doi.org/10.1021/ci300274c | J. Chem. Inf. Model. 2012, 52, 1806−1811

similar and that corresponding R-group positions were unambiguously assigned.

Activity Cliff Definition. Compounds comprising each target set were compared in a pairwise manner to identify activity cliffs. The following cliff criteria had to be met:

(i) A compound pair falls into structural category 1, 2, or 3.
(ii) The potency difference between compounds in a pair is at least 100-fold.

RESULTS AND DISCUSSION

Compound Data and Target Sets. From the current release of ChEMBL, a total of 33,746 compounds were assembled that were active against 542 human targets with 56,743 reported K_i measurements. On the basis of these data, 470 target sets containing multiple compounds were obtained for further analysis. These target sets contained variable numbers of compounds, ranging from two to 2058 (adenosine A2a receptor ligands). In addition, the target sets covered a wide range of target families including, among others, different protein kinases, various proteases, G-protein coupled receptors (GPCRs), and ion transporters.

Activity Cliff Criteria. For pairs of compounds forming activity cliffs, we required a potency difference of at least two orders to focus the analysis on large-magnitude cliffs, in accord with earlier investigations.[5,6] A major goal of our study has been the introduction of a well-defined structural framework for activity cliff formation, as an alternative to calculated similarity values or substructure relationships. Therefore, three structural categories were introduced on the basis of scaffold/R-group decomposition, as detailed in the Materials and Methods section. Compound pairs falling into each category were characterized by well-defined similarity relationships that did not require taking additional criteria (or descriptors) into account. Activity cliffs falling into each category were characterized by the presence of unique structural relationships. On the basis of this categorization, different classes/types of activity cliffs were defined, as specified below.

Activity Cliff Distribution. For each structural category, we first systematically identified all activity cliffs in ChEMBL target sets. The overall cliff distribution is reported in Table 1. A total of 10,589 category-based activity cliffs were detected in 235 target sets. These target sets contained from four to 2058 compounds. On average, each of these activity cliff-containing sets consisted of 223 compounds. A scatter plot reflecting the correlation between the number of compounds per target set and the number of the activity cliffs is shown in Figure 3. Only limited correlation was observed (correlation coefficient of 0.48). Furthermore, a total of 6200 compounds (i.e., ~19% of all active compounds subjected to our analysis) were involved in the formation of activity cliffs. These compounds represented 1269 unique scaffolds and 3718 unique R-group sets. For structural category 1, 2, and 3, we detected 362, 9597, and 630 cliffs, respectively (originating from 98, 230, and 109 different target sets, respectively). Hence, the majority of qualifying cliffs belonged to category 2, i.e., cliff forming compounds had identical scaffolds but different R-group sets. This was expected because this category covered many series of analogs.

Table 1. Activity Cliffs, Compounds, and Target Sets[a]

Structural category	Δ Size	# Cliffs	# Cpds	# Target sets	# Scaffolds	# R-groups
(1) same scaffold and same R-groups	0	362	552	98	183	201
(2) same scaffold and different R-groups	<3	9597	5387	220	716	3653
(3) different scaffolds and same R-groups	0	630	800	109	672	198
total	–	10,589	6200	235	1269	3718

[a]For each structural category, the size difference (Δ Size) of compounds forming a pair is reported (as the number of non-hydrogen atoms). The total number of activity cliffs and of compounds involved in cliff formation is given. In addition, numbers of target sets, scaffolds they contain, and unique R-group sets are provided.

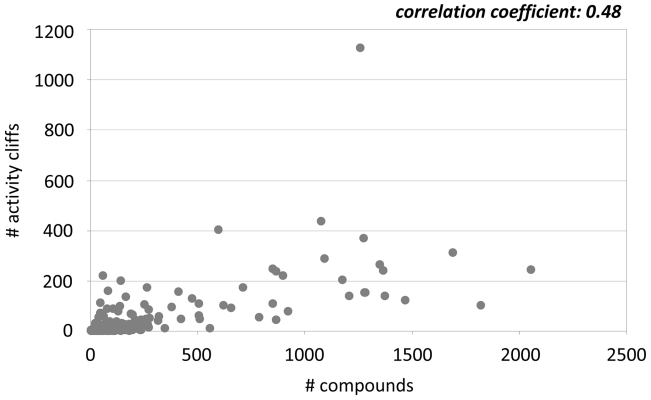


Figure 3. Compounds vs activity cliffs. The correlation between the number of compounds per target and the corresponding number of activity cliffs is reported in a scatter plot. Each data point represents a different target set.

Similarity Comparison. For all activity cliff pairs in category 1–3, Tanimoto similarity using MACCS structural keys was calculated, as reported in Figure 4a–c. Corresponding cumulative Tc distributions are shown in Figure 4d–f. Most activity cliffs in category 1 yielded very high Tanimoto coefficient values of >0.9 (Figure 4a), which was expected for cliff partners with conserved scaffolds and R-group sets. By contrast, similarity values of activity cliffs belonging to category 2 displayed a wide range of similarity values, despite the conservation of scaffolds and small size difference between cliff partners (Figure 4b). Thus, many of these cliffs would have remained undetected on the basis of Tanimoto similarity, although they were formed by analogs. This reflects a caveat of similarity calculations. A similar trend was observed for activity cliffs in category 3 that were characterized by the presence of different topologically equivalent scaffolds and shared R-group sets having either conserved or nonconserved topology (Figure 4c). We have previously shown that MACCS and other fingerprint types such as extended connectivity fingerprints[12] yield similar activity cliff distributions if Tc threshold values for cliff formation are appropriately set.[5] For example, a MACCS Tc value of 0.85 closely corresponds to a Tc value of 0.55 for extended connectivity fingerprint with bond diameter 4.[5]

Activity Cliff Classification. On the basis of our structural categorization, we then further refined the classification of activity cliffs and introduced five different types of cliffs, as reported in Table 2. Activity cliffs falling into category 1 were further divided into cliffs in which partners only differed by

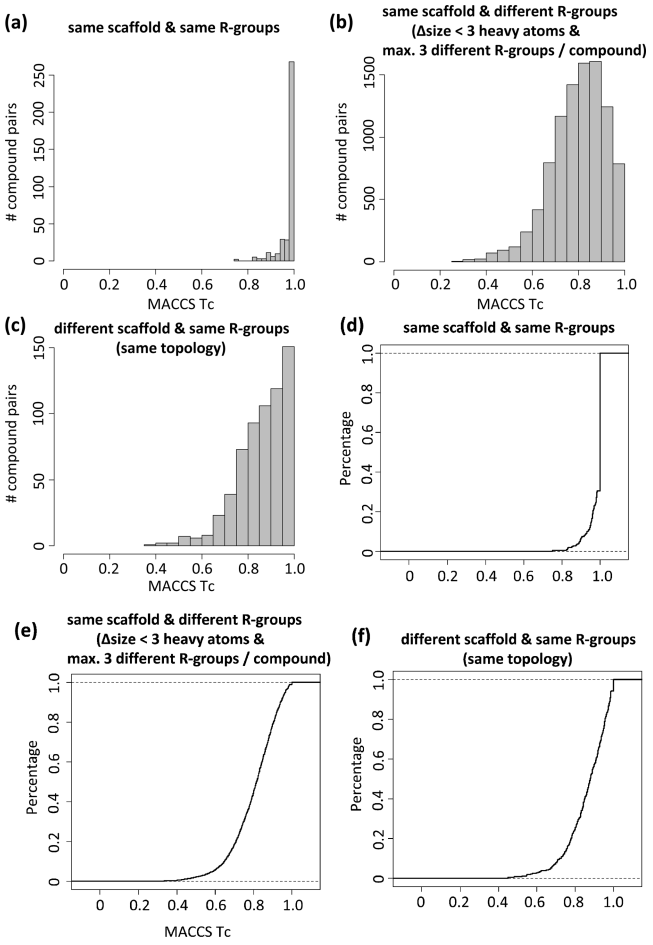


Figure 4. Tanimoto similarity. For activity cliffs falling into each structural category, the distribution of Tanimoto coefficient values (Tc) is reported in panels a–c. Corresponding cumulative Tc distributions are shown in panels d–f.

Table 2. Activity Cliff Classification[a]

Structural category	Activity cliff type	# Cliffs	Definition Cliff partners are distinguished by:
(1) same scaffold and same R-groups	chirality cliffs	162	chirality only (R-group topology is conserved)
	topology cliffs	200	topology of conserved R-groups
(2) same scaffold and different R-groups	R-group cliffs	9597	composition of R-group sets
(3) different scaffolds and same R-groups	scaffold cliffs	468	topologically equivalent scaffolds (R-group topology is conserved)
	scaffold/ topology cliffs	162	topologically equivalent scaffolds and R-group topology

[a]Five different types of activity cliffs are defined. For each type, the total number of activity cliffs found in ChEMBL is reported.

chirality at one or more centers and cliffs with different R-group topology (i.e., with identical R-group sets but at least partly noncorresponding R-group positions). These activity cliff types were termed "chirality cliffs" and "topology cliffs", respectively. Both types of activity cliffs are chemically intuitive. It is well appreciated that differences in stereochemistry are activity determinants in many instances. Accordingly, a considerable number of chirality cliffs have been identified in our analysis. In addition, a comparable number of topology cliffs has been detected (see below). Topology cliffs are chemically relevant because in this case, positional effects of conserved R-groups lead to cliff formation. Hence, the presence of topology cliffs indicates that the relative orientation of substitution site(s), rather than the chemical nature of substituents, is an important SAR determinant.

All activity cliffs in category 2 only differed by R-group content and were thus designated "R-group cliffs". These activity cliffs comprise the largest group and would also be obtained if a scaffold-centric analysis would be carried out, for example, on the basis of R-group decomposition. It should also be noted that more stringent criteria were applied to further ensure the structural similarity of compounds sharing the same scaffold, i.e., the size difference of R-groups was limited to less than three non-hydrogen atoms and each compound forming a pair was permitted to contain at most three R-groups. Moreover, other structurally well-defined types of activity cliffs would not be identified.

Activity cliffs falling into category 3 were also divided into two types including "scaffold cliffs" where cliff partners were only distinguished by topologically equivalent scaffolds and, in addition, "scaffold/topology cliffs" in which R-group topology also differed.

Table 2 reports the frequency of occurrence of these five different types of activity cliffs over all target sets. We detected 162 chirality cliffs, 200 topology cliffs, 9597 R-group cliffs, 468 scaffold cliffs, and 162 scaffold/topology cliffs. Thus, four of five cliff types occurred with similar frequency, with the exception of R-group cliffs for which 9597 instances were found (and that were thus on average ~40 times more frequent than the other four cliff types). This was expected because R-group cliffs covered activity cliffs in analog series (see above). These series typically originate from chemical optimization efforts and represent the major part of compound data from medicinal chemistry sources.

Representative Activity Cliffs. In Figure 5, examples of activity cliffs are shown for each of the five different classes. Similar to earlier observations,[6,7] we also found that many newly defined activity cliffs were coordinated, i.e., subsets of active compounds formed overlapping cliffs, as illustrated in Figure 5. A notable exception was the class of topology cliffs where many instances of single cliffs formed by individual pairs of compounds were detected. Figure 5a shows typical examples of chirality cliffs, formed by compounds active against the mu opioid receptor. In this case, stereochemical differences at several atoms of a central aliphatic ring moiety led to potency variations of more than 2 orders of magnitude. Figure 5b shows topology cliffs formed by delta opioid receptor ligands. Here, the position of a single hydroxyl group varied among mostly highly potent compounds. However, the presence of the hydroxyl group at the ortho position of the phenyl moiety at the lower left led to a dramatic reduction in potency of nearly 3 orders of magnitude. Figure 5c shows R-group cliffs formed by compounds active against the nociceptin receptor. This series is based upon a scaffold that was a notable center of activity cliff formation. Compounds represented by this scaffold formed a total of 311 cliffs within this target set. Such examples also provide a rationale for the high frequency of occurrence of R-group cliffs compared to other cliff types. Single or multiple R-group replacements led to potency variations of nearly 4 orders of magnitude within this analog series. In Figure 5d, scaffold cliffs are shown that were formed by coagulation factor X inhibitors. This is another interesting case where a conserved

dx.doi.org/10.1021/ci300274c | J. Chem. Inf. Model. 2012, 52, 1806–1811

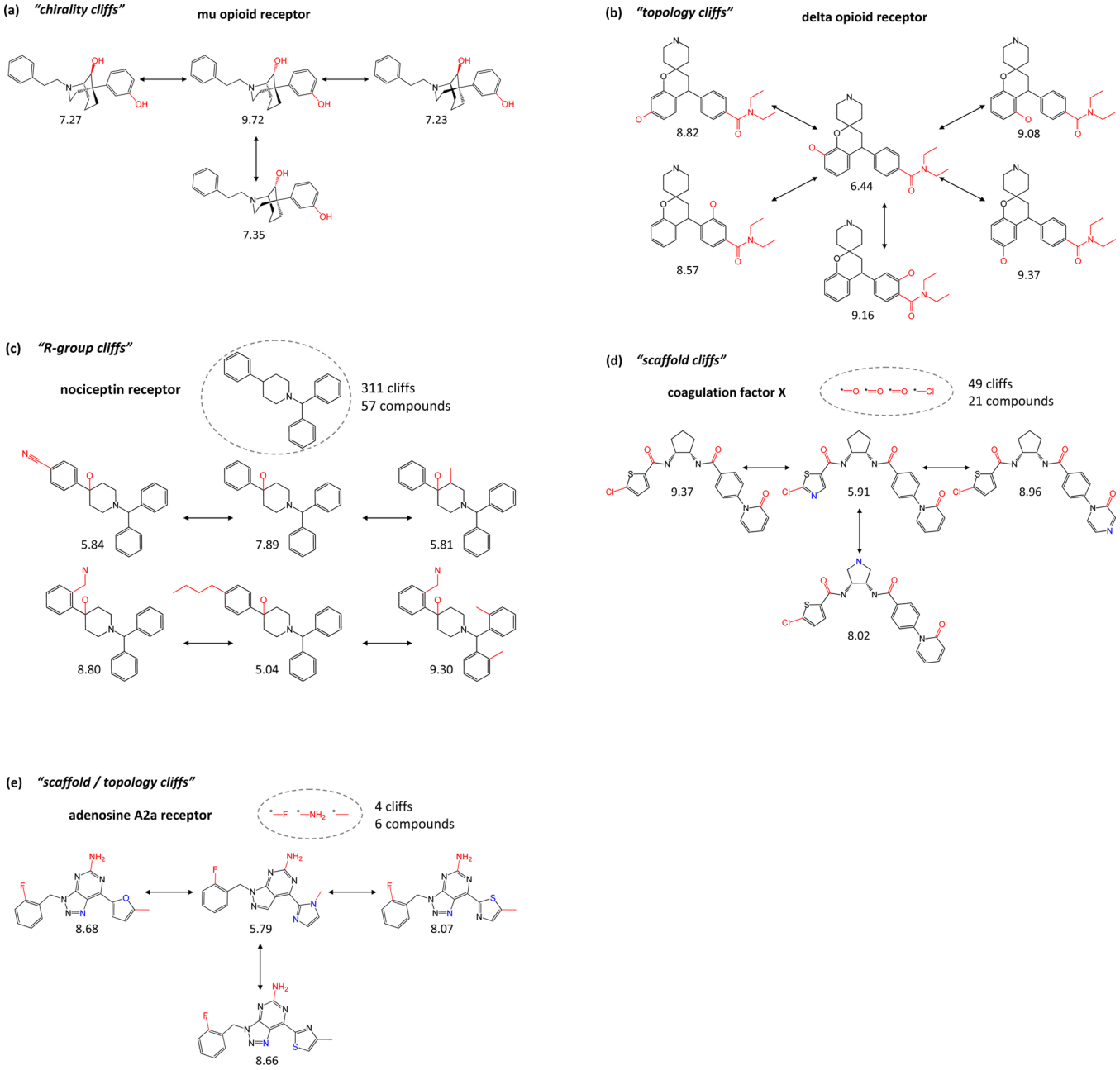


Figure 5. Representative activity cliffs. For each activity cliff type, representative examples are provided. (a) Shown are three "chirality cliffs" involving four compounds active against the mu opioid receptor. The potency of each compound is reported (pK_i values). (b) Five "topology cliffs" are shown that were formed by six compounds active against the delta opioid receptor. These compounds shared the same scaffold and R-group set. However, differences in the position of the hydroxyl group led to significant potency alterations. (c) At the top, a four-ring scaffold is displayed that represented 57 compounds involved in the formation of 311 activity cliffs. Six "R-group cliffs" are shown involving six compounds active against the nociceptin receptor. (d) Three "scaffold cliffs" were formed by four compounds active against coagulation factor X. These topologically equivalent scaffolds only differed by the position of a single nitrogen atom. (e) Three "scaffold/topology cliffs" are shown involving four adenosine A2a receptor ligands that comprised topologically equivalent scaffolds and a conserved R-group set. The only difference in R-group topology was the change of a methyl position from ortho to meta.

set of four R-groups attached to different scaffolds with constant topology occurred in 21 compounds forming 49 cliffs. The four-ring scaffolds in Figure 5d are closely related and only distinguished by a single nitrogen atom. However, these minute differences caused the formation of scaffold cliffs. Finally, in Figure 5e, examples of scaffold/topology cliffs among adenosine A2a receptor ligands are shown that involved heteroatom changes in scaffolds and topology variations among a set of three R-groups.

Concluding Remarks. The assessment of activity cliffs stringently depends on applied potency and similarity criteria, which, by necessity, are often subjectively chosen. Generally accepted criteria for the formation of activity cliffs and general classification schemes are currently unavailable. The way

dx.doi.org/10.1021/ci300274c | J. Chem. Inf. Model. 2012, 52, 1806−1811

molecular similarity is assessed is the most critical variable for activity cliff assessment. Similarity relationships that are calculated on the basis of molecular descriptors are often difficult to interpret in chemical terms and complicate activity cliff analysis. Also, within analog series, chemically intuitive activity cliffs might not be detected if calculated similarity values fall below prespecified thresholds, which we often observe, depending on the specific features of compound classes and chosen representations. In light of this situation, we have developed a new structural classification scheme for activity cliffs, leading to the introduction of five different classes of cliffs. In the ChEMBL compound collection, all five types of activity cliffs frequently occur. This has been shown by extracting all currently available cliffs spanning at least 2 orders of magnitude in potency from ChEMBL compounds. Importantly, the five types of cliffs are distinguished by local structural changes within different structural contexts leading potency variations of large magnitude. Thus, this classification goes beyond the assessment of substructure relationships and provides a further refined view of activity cliffs. Chirality, topology, scaffold, and R-group cliffs are characterized by the presence of well-defined and limited structural differences that are readily interpretable from a chemical perspective. These cliffs capture SAR information as we have illustrated for series of compounds forming activity cliffs of different types. We anticipate that compound structure- and topology-based organization of activity cliffs will further increase their attractiveness for SAR data mining and analysis. In medicinal chemistry, activity cliffs identified in individual compound series or larger data sets are typically initial focal points of SAR exploration because they often provide immediate insights into SAR critical substitution sites, R-groups, and/or core structure features. Such insights might then be directly translated into the design of new analogs or the selection of alternative scaffolds. For these purposes, the structural categorization of activity cliffs introduced herein should be particularly helpful because it strongly supports the chemical interpretation of compound modifications leading to large potency effects.

Upon publication of this article, the complete compendium of ChEMBL activity cliffs belonging to the five newly introduced classes can be obtained without restriction via the following URL: http://www.lifescienceinformatics.uni-bonn.de.

AUTHOR INFORMATION

Corresponding Author

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

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Maggiora, G. M. On outliers and activity cliffs – Why QSAR often disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535–1535.

(2) Bajorath, J.; Peltason, L.; Wawer, M.; Guha, R.; Lajiness, M. S.; van Drie, J. H. Navigating structure–activity landscapes. *Drug Discovery Today* **2009**, *14*, 698–705.

(3) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity landscape representations for structure–activity relationship analysis. *J. Med. Chem.* **2010**, *53*, 8209–8223.

(4) Stumpfe, D.; Bajorath, J. Exploring activity cliffs in medicinal chemistry. *J. Med. Chem.* **2012**, *55*, 2932–2942.

(5) Wassermann, A. M.; Dimova, D.; Bajorath, J. Comprehensive analysis of single- and multi-target activity cliffs formed by currently available bioactive compounds. *Chem. Biol. Drug Des.* **2011**, *78*, 224–228.

(6) Vogt, M.; Huang, Y.; Bajorath, J. From activity cliffs to activity ridges: Informative data structures for SAR analysis. *J. Chem. Inf. Model.* **2011**, *51*, 1848–1856.

(7) Namasivayam, V.; Bajorath, J. Searching for coordinated activity cliffs using particle swarm optimization. *J. Chem. Inf. Model.* **2012**, *52*, 927–934.

(8) Guha, R.; Van Drie, J. H. Structure–activity landscape index: Identifying and quantifying activity cliffs. *J. Chem. Inf. Model.* **2008**, *48*, 646–658.

(9) Stumpfe, D.; Bajorath, J. Assessing the confidence level of public domain compound activity data and the impact of alternative potency measurements on SAR analysis. *J. Chem. Inf. Model.* **2011**, *51*, 3131–3137.

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

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

(12) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(13) Hu, Y.; Bajorath, J. Exploration of 3D activity cliffs on the basis of compound binding modes and comparison of 2D and 3D cliffs. *J. Chem. Inf. Model.* **2012**, *52*, 670–677.

(14) Peltason, L.; Bajorath, J. Molecular similarity analysis uncovers the presence of heterogeneous structure–activity relationships and variable activity landscapes within active sites. *Chem. Biol.* **2007**, *14*, 489–497.

(15) Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2004; pp 271–285.

(16) Wassermann, A. M.; Bajorath, J. Chemical substitutions that introduce activity cliffs across different compound classes and biological targets. *J. Chem. Inf. Model.* **2010**, *50*, 1248–1256.

(17) Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J. MMP-Cliffs: Systematic Iientification of activity cliffs on the basis of matched molecular pairs. *J. Chem. Inf. Model.* **2012**, *52*, 1138–1145.

(18) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–D1107.

(19) *ChEMBL*. http://www.ebi.ac.uk/chembldb/ (accessed May 8, 2012).

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

(21) Xu, Y.-J.; Johnson, M. Algorithm for naming molecular equivalence classes represented by labeled pseudographs. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 181–185.