Rationalizing the Formation of Activity Cliffs in Different Compound Data Sets

Huabin Hu, Dagmar Stumpfe, and Jürgen Bajorath*  
Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Endenicher Allee 19c, D-53115 Bonn, Germany

ABSTRACT: Activity cliffs are formed by structurally analogous compounds with large potency variations and are highly relevant for the exploration of discontinuous structure–activity relationships and compound optimization. So far, activity cliffs have mostly been studied on a case-by-case basis or assessed by global statistical analysis. Different from previous investigations, we report a large-scale analysis of activity cliff formation with a strong focus on individual compound activity classes (target sets). Compound potency distributions were systematically analyzed and categorized, and structural relationships were dissected and visualized on a per-set basis. Our study uncovered target set-dependent interplay of potency distributions and structural relationships and revealed the presence of activity cliffs and origins of cliff formation in different structure–activity relationship environments.

1. INTRODUCTION  
Activity cliffs are formed by structurally similar (analogous) active compounds with large differences in potency. Because activity cliffs represent small chemical changes having large biological activity effects, they embody the pinnacle of structure–activity relationship (SAR) discontinuity, which is detrimental for quantitative SAR predictions. However, discontinuous SARs and activity cliffs often reveal SAR determinants, especially when encountered during early stages of compound optimization, and thus provide viable information for medicinal chemistry.

For a consistent assessment of activity cliffs, similarity and potency difference criteria must be clearly defined. On the basis of globally assessed potency range distributions of pairs of active analogues, an at least 100-fold difference in potency (on the basis of equilibrium constants, if available) has been proposed and frequently used as an activity cliff criterion. The definition of activity cliffs also depends on the molecular representations and similarity measures that are used. Compound similarity for activity cliff definition can be quantified in different ways, for example, by calculating Tanimoto similarity on the basis of molecular fingerprint representations or by applying substructure-based similarity criteria. Numerical similarity measures, such as the Tanimoto coefficient, yield a continuum of values, and a threshold must be set for defining activity cliffs. By contrast, substructure-based methods produce a binary readout, for example, two compounds share the same core structure—and are classified as similar—or they do not. In addition to comparing molecular graph-based (two-dimensional) representations, activity cliffs have also been determined in three dimensions by calculating the similarity of experimental compound binding modes taken from complex X-ray structures.

For graph-based activity cliff definition, substructure similarity assessment is—in our experience—generally more consistent than numerical similarity calculations and often easier to interpret from a chemical perspective. Among substructure-based approaches, the matched molecular pair (MMP) concept is particularly attractive for activity cliff definition. An MMP is defined as a pair of compounds that are only distinguished by a chemical modification at a single site. This modification corresponds to the exchange of a pair of substructures, which is termed a chemical transformation. By introducing appropriate transformation size restrictions, the formation of MMPs can be limited to structural analogues typically generated during compound optimization. Applying this similarity criterion yields a structurally conservative and chemically intuitive definition of activity cliffs. Moreover, transformation size-restricted MMPs can be efficiently generated algorithmically, hence enabling large-scale analysis of activity cliff populations.

In light of these considerations, our preferred activity cliff definition encompasses the formation of a transformation size-restricted MMP by two compounds sharing the same biological activity that have an at least 100-fold difference in potency. Whenever possible, potency differences are determined on the basis of (assay-independent) equilibrium constants. The so-defined activity cliffs have been termed MMP-cliffs.

The definition of activity cliffs is focused on compound pairs and hence accounts for pairwise relationships. However, activity cliffs in compound data sets are mostly not formed by isolated compound pairs (i.e., pairs without structural neighbors forming...
additional activity cliffs). Rather, the vast majority of activity cliffs are formed in a coordinated manner by groups of structurally related compounds with large potency variations, meaning that individual compounds are involved in the formation of multiple activity cliffs with different analogues. 

In activity cliff networks where nodes represent compounds and edges pairwise activity cliffs, compound subsets forming coordinated cliffs give rise to the formation of disjoint clusters. These activity cliff clusters are a rich source of SAR information and much more informative than cliffs considered as isolated. More than 95% of MMP-cliffs detected across different data sets were formed in a coordinated manner. In activity cliff networks, clusters often include “hubs,” that are, nodes representing molecules that are centers of local activity cliff formation with multiple partner compounds. Such molecules have also been termed “activity cliff generators.”

In addition to activity cliff coordination, the frequency with which activity cliffs occur across different data sets has been determined. There has been substantial growth in activity cliff information over time. For example, from June 2011 until January 2015, the number of MMP-cliffs originating from the ChEMBL database, the major public repository of compounds and activity data from medicinal chemistry sources, nearly doubled; with a total of more than 17 000 MMP-cliffs available at the beginning of 2015. In addition, the target coverage of MMP-cliffs increased from about 200 to 300 individual target proteins over this period of time. However, despite this strong growth, the proportion of bioactive compounds involved in the formation of MMP-cliffs across different compound data sets remained essentially constant at close to 23%.

So far, activity cliffs have been studied in exemplary compound sets on a case-by-case basis or surveyed by global statistical analysis. In addition, cliff populations have been organized and visualized in network representations. However, what has not been attempted thus far is systematically exploring and comparing activity cliff formation in different compound activity classes (also called target sets). To these ends, we have analyzed in detail potency distributions and structural relationships between compounds in many different target sets, studied how activity cliffs were formed, and determined the differences between sets. Hence, the focus of our current study has been on details of activity cliff arrangements in individual compound sets rather than on global statistical exploration. Our analysis revealed many characteristic differences in activity cliff formation between target sets.

Figure 1. Potency distribution in target sets and categorization. The compound potency distributions of all 237 target sets were analyzed in a boxplot and the IQR, that is, the difference between quartile 3 and 1, was determined. On the basis of the IQR, target sets were divided into three different categories (CAT 1: IQR < 1; CAT 2: 1 ≤ IQR < 2; CAT 3: IQR ≥ 2).

2. MATERIALS AND METHODS

2.1. Activity Cliff Definition. For our current analysis, we introduced a modification of our preferred MMP-cliff definition stated above. For MMP generation, standard random fragmentation of exocyclic single bonds was replaced by fragmentation according to retrosynthetic (RECAP) rules, yielding (transformation size-restricted) RECAP-MMPs (RMMPs). Retrosynthetic MMPs were generated to further increase the chemical relevance (synthetic accessibility) of compound pairs, forming cliffs. Accordingly, the formation of an RMMP was used as a similarity criterion for activity cliffs, and an at least 100-fold difference in potency between RMMP compounds was required, as before. The so-defined activity cliffs are referred to as RMMP-cliffs.

2.2. Compounds and Activity Data. Bioactive compounds with high-confidence activity data were assembled from ChEMBL version 23. The following selection criteria were applied: First, only compounds involved in direct interactions (type “D”) with human targets at the highest confidence level (assay confidence score 9) were selected. Second, only numerically specified equilibrium constants ($K_i$ values) were considered as potency measurements. Equilibrium constants were reported as p$K_i$ values. On the basis of these selection criteria, a total of 71 967 unique compounds were obtained with activity against a total of 904 targets. Accordingly, these compounds were organized into 904 target sets.

2.3. RMMP Analysis. RMMPs were systematically generated for all target sets, yielding 354 094 target set-based RMMPs (243 110 unique RMMPs) that were formed by 46 977 compounds from 574 target sets. For the subsequent analysis, only target sets that contained at least 100 RMMPs were retained, which resulted in 237 sets yielding a total of 347 025 target-based RMMPs (238 795 unique RMMPs) formed by 44 451 compounds.

For each target set, an RMMP network was generated in which nodes represented compounds and edges pairwise RMMP relationships. In this network, each separate RMMP cluster represented a unique series of analogues. RMMP networks were also used to represent RMMP-cliffs by highlighting edges that represented both RMMP and activity cliff relationships. All network representations were drawn with Cytoscape.

2.4. Potency Distributions. For the 237 qualifying target sets, compound potency distributions were monitored in boxplots. On the basis of the interquartile range (IQR), that is, the range between quartile 1 (Q1) and 3 (Q3), target sets...
we were assigned to three different categories, as shown in Figure 1: category 1 (CAT 1), IQR was smaller than 1 order of magnitude (<10-fold difference in potency); CAT 2, IQR fell between 1 and less than 2 orders of magnitude (10- to 100-fold difference); and CAT 3, IQR no smaller than 2 orders of magnitude (≥100-fold difference in potency). By definition, the IQR represented the potency range of ~50% of the compounds in each target set.

3. RESULTS AND DISCUSSION

3.1. Study Concept. Activity cliffs have so far mostly been studied on the basis of individual compound series or by global statistical analysis.3−5 Our current study was designed to systematically investigate, for the first time, the differences in activity cliff formation and frequency between different target sets by relating compound potency distributions and structural relationships to each other. Therefore, potency distributions were determined for many different target sets, categorized, compared, and related to intra-set analogue relationships, which were systematically determined. Primary goals of the analysis included the assessment of differences in activity cliff formation and frequency between different target sets and the rationalization of such differences on the basis of potency and structural criteria, as defined in the following. To better understand target set-dependent activity cliff distributions, they were visualized in network representations. Taken together, these features set our current analysis apart from previous studies of activity cliffs in computational and medicinal chemistry.3,4

3.2. Structural Relationships. Close structural relationships between active compounds are one of the two major determinants of activity cliffs, in addition to potency differences. RMMP (or MMP) calculations reveal close structural relationships and identify pairs of analogues. Importantly, however, the number of RMMPs produced by a given target set cannot be reliably used as an indicator of structural homogeneity. Rather, the presence or absence of multiple subsets of analogues comprising different series strongly influences structural heterogeneity or homogeneity, which is reflected by the cluster structure of RMMP networks, as illustrated in Figure 2. Here, two target sets with similar numbers of RMMP-forming compounds are compared. The target set on the left was dominated by a large cluster of analogues and was thus structurally homogeneous, whereas the set on the right contained 20 different small clusters and 1 larger cluster and was structurally heterogeneous. It follows that the cluster structure of RMMP networks must be carefully considered as a prerequisite for RMMP-cliff formation.

3.3. Potency Distributions and Profiles. The likelihood of large potency differences between similar compounds can be estimated by monitoring the potency distributions of target sets. For our analysis, we assigned potency distributions to three different categories (CAT 1−3) on the basis of boxplot-derived IQR values, as shown in Figure 1. CAT 1, 2, and 3 comprised 25, 169, and 43 target sets, respectively. Hence, the majority of target sets fell into CAT 2 whose IQR spanned 1 to 2 orders of magnitude in potency and thus delineated an activity cliff-relevant range, which was further expanded by CAT 3. These observations supported our categorization of potency distributions. Accordingly, potency distributions became increasingly variable from CAT 1 to 3, as revealed by the potency distribution profiles in Figure 3. The CAT 1 profiles in Figure 3 reflect narrow potency distributions on the basis of which activity cliff formation is unlikely. By contrast, the CAT 2 profiles in Figure 3b and, especially, CAT 3 profiles in Figure 3c reveal large potency variations between structural analogues, resulting in a principally high propensity of activity cliffs.

3.4. RMMP-Cliffs. In 207 of the 237 qualifying targets sets, RMMP-cliffs were identified, amounting to a total of 11 834 cliffs. Table 1 reports that the number of RMMP-cliffs increased over target sets of CAT 1, 2, and 3, with on average 2, 52, and 69 cliffs per set, respectively. Thus, there was a general trend of increasing number of RMMP-cliffs with increasing variability of potency distributions. The very small number of RMMP-cliffs for CAT 1 sets was directly attributable to the narrow potency distributions characterizing this category. Table 2 reports that the 48 target sets containing 50 to a maximum of 820 RMMP-cliffs exclusively belonged to CAT 2 and 3 that had activity cliff-relevant IQR values. By contrast, target sets with less than 50 RMMP-cliffs were found in all 3 categories. Figure 4 shows that the majority of target sets with large number of 100 or more
RMMP-cliffs belonged to CAT 2, which was due to the large number of 169 target sets in this category compared to only 43 sets in CAT 3. A systematic increase in the number of activity cliffs with increasing IQR values was not observed. However, despite these general trends, the propensity to form RMMP-cliffs could not solely be attributed to the variability and spread of potency distributions. Rather, as further discussed below, potency distributions in target sets must be viewed in combination with RMMP networks and their cluster structure. Table 1 also reports that target sets in CAT 1, 2, and 3 contained on average 10, 54, and 37 RMMP clusters, respectively. Thus, CAT 2 and CAT 3 sets contained large number of clusters (analogue series) whose local potency distributions strongly influenced RMMP-cliff formation.

### 3.5. Interplay of Potency Patterns and Structural Relationships

The 207 target sets containing RMMP-cliffs were individually examined to evaluate potency distribution profiles and RMMP networks in context and rationalize why RMMP-cliffs were formed with different frequencies. The analysis revealed a number of characteristic features determining cliff formation that are summarized in Figure 5 by comparing exemplary target sets.

**Figure 3.** Potency distribution profiles. Shown are exemplary potency distribution profiles for target sets belonging to different categories [(a), CAT 1; (b), CAT 2; (c), CAT 3] according to Figure 1. Black dots represent RMMP compounds and red dots singletons not participating in RMMPs.

**Table 1. Target Set Statistics**

| CAT | # target sets | # clusters (mean) | # RMMP-cliffs (mean) |
|-----|---------------|-------------------|----------------------|
| 1   | 25            | 10                | 2                    |
| 2   | 169           | 54                | 52                   |
| 3   | 43            | 37                | 69                   |

For each target set category (CAT), the number (#) of target sets, mean number of RMMP clusters per set, and mean number of RMMP-cliffs are reported.

**Table 2. RMMP-Cliff Distribution**

| # RMMP-cliffs (range) | # target sets | CATs |
|-----------------------|---------------|------|
| (0, 10)               | 77            | 1, 2, 3 |
| (10, 20)              | 33            | 1, 2, 3 |
| (20, 50)              | 49            | 1, 2, 3 |
| (50, 100)             | 20            | 2, 3  |
| (100, 500)            | 25            | 2, 3  |
| (500, 820)            | 3             | 2, 3  |

For different ranges of RMMP-cliffs, the number of target sets (# targets) and categories (CATs) they belong to are reported.

**Figure 4.** RMMP-cliffs vs IQR values. For each of the 237 target sets, the number of RMMP-cliffs (y-axis) is plotted against increasing IQR values (x-axis). Red vertical lines separate target sets belonging to CAT 1, 2, and 3.
Figure 5. continued
Figure 5. continued
extensive RMMP relationships. In addition, Figure 5a (bottom) displays somatostatin receptor 5 ligands with a variable CAT 2 distribution and more than 100 RMMP-forming compounds. Although cliff formation was more likely in this case, the target set did not contain any RMMP-cliffs either. This was a direct consequence of a heterogeneous cluster structure and local

Figure 5. Differences in RMMP-cliff formation. In (a−c), exemplary target sets with characteristic differences in activity cliff formation are compared, as described in the text. For each set, its potency distribution profile and RMMP network are shown and RMMP statistics are reported. Network nodes are colored by potency using a continuous color spectrum from red (lowest potency in the target set) over yellow (intermediate) to green (highest potency). If available, compounds forming exemplary RMMP-cliffs are shown and consistently labeled in all display items.
potency distributions over different subsets of analogues forming 16 clusters, as revealed by the RMMP network of this set.

Figure 5b shows two different sets of kinase inhibitors with similar CAT 2 potency distributions but different RMMP cluster structures that yielded 40 (top) and 27 (bottom) RMMP-cliffs, respectively. Exemplary RMMP-cliffs are displayed in both instances, the target sets were structurally heterogeneous but RMMP-cliffs were formed across different clusters, revealing high degrees of SAR discontinuity.

In Figure 5c, sets of anandamide amidohydrolase (top) and Bcl-X (bottom) inhibitors are compared having CAT 2 (top) and CAT 3 (bottom) distributions, respectively. The anandamide amidohydrolase inhibitors contained only 49 RMMP-forming compounds. The RMMP network was dominated by a densely connected cluster of 19 analogues that formed 79 coordinated RMMP-cliffs (exemplary cliffs are shown). Thus, in this case, the number of RMMP-cliffs was much larger than the number of participating analogues because of extensive coordination of cliffs. Hence, this cluster represented an SAR hotspot. By contrast, the Bcl-X inhibitors contained a much larger number of 119 RMMP-forming compounds that were distributed over 20 clusters. Although the CAT 3 potency distribution of this target set was highly variable, the majority of compounds in individual clusters had comparable potency, whereas the potency levels of clusters significantly differed, giving rise to the presence of only three RMMP-cliffs.

Taken together, the results in Figure 5 were representative of many target sets we studied. Analyzing the potency distribution profiles and in combination with RMMP networks revealed the characteristic features of target sets and clearly rationalized differences in RMMP-cliff frequency across target sets.

4. CONCLUSIONS
Herein, we have reported a systematic analysis of RMMP-cliffs in more than 200 target sets to investigate and better understand the origins of cliff formation and differences in the frequency of cliffs. Our study was strongly focused on individual target sets and their comparison. Potency distributions were determined and categorized, and structural relationships were analyzed at the level of RMMPs and organized in networks. Structural homogeneity of target sets and potency distributions of increasing variability generally supported the formation of RMMP-cliffs. However, the interplay of structural and potency relationships determined the frequency with which RMMP-cliffs were formed, as revealed by relating potency profiles and RMMP networks to each other and studying local potency distributions across different RMMP clusters.

The analysis scheme introduced herein reveals target set-dependent formation of activity cliffs, provides immediate visual access to characteristic activity cliff-relevant features of target sets, and rationalizes differences in the frequency of cliffs across sets.

Author Contributions
The study was carried out by all authors, and the manuscript was written with contributions of all authors. All authors have approved the final version of the manuscript.

Notes
The authors declare no competing financial interest.

ACKNOWLEDGMENTS
H.H. is supported by the China Scholarship Council (CSC).

REFERENCES
(1) Lajiness, M. Evaluation of the Performance of Dissimilarity Selection Methodology. In QSAR: Rational Approaches to the Design of Bioactive Compounds; Silipo, C., Vittoria, A., Eds.; Elsevier: Amsterdam, Netherlands, 1991; pp 201–204.
(2) Maggiora, G. M. On Outliers and Activity Cliffs — Why QSAR often Disappoints. J. Chem. Inf. Model. 2006, 46, 1535.
(3) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. J. Med. Chem. 2012, SS, 2932–2942.
(4) Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. J. Med. Chem. 2014, 57, 18–28.
(5) Stumpfe, D.; Bajorath, J. Frequency of Occurrence and Potency Range Distribution of Activity Cliffs in Bioactive Cliffs. J. Chem. Inf. Model. 2012, 52, 2348–2353.
(6) Medina-Franco, J. L.; Martínez-Mayorga, K.; Bender, A.; Marín, R. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. Characterization of Activity Landscapes using 2D and 3D Similarity Methods: Consensus Activity Cliffs. J. Chem. Inf. Model. 2009, 49, 477–491.
(7) Hu, Y.; Furtmann, N.; Gütschow, M.; Bajorath, J. Systematic Identification and Classification of Three-Dimensional Activity Cliffs. J. Chem. Inf. Model. 2012, 52, 1490–1498.
(8) Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In Chemoinformatics in Drug Discovery; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2005; pp 271–285.
(9) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. J. Chem. Inf. Model. 2010, 50, 339–348.
(10) Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J. MMP-Cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs. J. Chem. Inf. Model. 2012, 52, 1138–1145.
(11) Vogt, M.; Huang, Y.; and Bajorath, J. From Activity Cliffs to Activity Ridges: Informative Data Structures for SAR Analysis. J. Chem. Inf. Model. 2011, 51, 1848–1856.
(12) Stumpfe, D.; Dimova, D.; Bajorath, J. Composition and Topology of Activity Cliff Clusters Formed by Bioactive Compounds. J. Chem. Inf. Model. 2014, 54, 451–461.
(13) Dimova, D.; Stumpfe, D.; Hu, Y.; Bajorath, J. Activity Cliff Clusters as a Source of Structure-Activity Relationship Information. Expert Opin. Drug Discovery 2015, 10, 441–447.
(14) Stumpfe, D.; Bajorath, J. Monitoring Global Growth of Activity Cliff Information over Time and Assessing Activity Cliff Frequencies and Distributions. Future Med. Chem. 2015, 7, 1565–1579.
(15) Méndez-Lucio, O.; Pérez-Villanueva, J.; Castillo, R.; Medina-Franco, J. L. Identifying Activity Cliff Generators of PPAR Ligands Using SAS Maps. Mol. Inf. 2012, 31, 837–846.
(16) Pérez-Villanueva, J.; Méndez-Lucio, O.; Soria-Arteche, O.; Medina-Franco, J. L. Activity Cliffs and Activity Cliff Generators Based on Chemotype-Related Activity Landscapes. Mol. Diversity 2015, 19, 1021–1035.
(17) 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.
(18) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP-Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful
Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* 1998, 38, 511–522.

(19) de la Vega de León, A.; Bajorath, J. Matched Molecular Pairs Derived by Retrosynthetic Fragmentation. *Med. Chem. Commun.* 2014, 5, 64–67.

(20) Smoot, M. E.; Ono, K.; Ruscheinski, J.; Wang, P.-L.; Ideker, T. Cytoscape 2.8: New Features for Data Integration and Network Visualization. *Bioinformatics* 2010, 27, 431–432.