JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

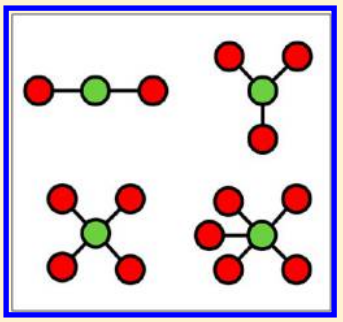
Composition and Topology of Activity Cliff Clusters Formed by Bioactive Compounds

Dagmar Stumpfe, Dilyana Dimova, 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

Supporting Information

ABSTRACT: The assessment of activity cliffs has thus far mostly focused on compound pairs, although the majority of activity cliffs are not formed in isolation but in a coordinated manner involving multiple active compounds and cliffs. However, the composition of coordinated activity cliff configurations and their topologies are unknown. Therefore, we have identified all activity cliff configurations formed by currently available bioactive compounds and analyzed them in network representations where activity cliff configurations occur as clusters. The composition, topology, frequency of occurrence, and target distribution of activity cliff clusters have been determined. A limited number of large cliff clusters with unique topologies were identified that were centers of activity cliff formation. These clusters originated from a small number of target sets. However, most clusters were of small to moderate size. Three basic topologies were sufficient to describe recurrent activity cliff cluster motifs/topologies. For example, frequently occurring clusters with star topology determined the scale-free character of the global activity cliff network and represented a characteristic activity cliff configuration. Large clusters with complex topology were often found to contain different combinations of basic topologies. Our study provides a first view of activity cliff configurations formed by currently available bioactive compounds and of the recurrent topologies of activity cliff clusters. Activity cliff clusters of defined topology can be selected, and from compounds forming the clusters, SAR information can be obtained. The SAR information of activity cliff clusters sharing a/one specific activity and topology can be compared.



INTRODUCTION

The activity cliff concept[1−4] has experienced increasing attention in computational and medicinal chemistry.[2−4] Originally, activity cliffs were defined as pairs of structurally similar active compounds having a large difference in potency.[1,2] Given this definition, the specification of similarity and potency difference criteria is of critical relevance for the assessment of activity cliffs.[2−4] The popularity of the activity cliff concept in medicinal chemistry is primarily due to the underlying "small chemical changes−large biological effects" paradigm, which assigns high structure−activity relationship (SAR) information content to activity cliffs.[2,3] In addition to SAR exploration, activity cliffs are of interest for computational analysis because they can be explored through systematic mining of compound activity data[3,4] and because they are focal points of activity landscape modeling.[5,6] A variety of molecular representations have been utilized to asses compound similarity in the analysis of activity cliffs, typically in combination with Tanimoto similarity calculations.[2,3,7] However, in medicinal chemistry, activity cliffs defined on the basis of such whole-molecule similarity calculations are often difficult to interpret.[3] Therefore, activity cliffs have also been defined on the basis of substructure relationships between active compounds,[3] for example, by employing the matched molecular pair (MMP) formalism.[8] An MMP is generally defined as a pair of compounds that only differ by a structural change at a single site,[8,9] i.e., the exchange of two substructures, a so-called chemical transformation.[9] By introducing transformation size restrictions,[10] such structural changes can be limited to small and chemically meaningful replacements that relate analogous compounds to each other. The formation of such transformation size-restricted MMPs has been applied as a similarity criterion for activity cliffs, leading to the introduction of MMP-cliffs.[10] The potency difference criterion is also critical for activity cliff analysis. Given the high relevance of activity cliffs for SAR analysis, the exclusive consideration of high-confidence activity data is strongly recommended.[3,4] A generally preferred activity cliff definition has been put forward that requires the formation of a transformation size-restricted MMP for cliff partners and the presence of a potency difference of at least 2 orders of magnitude on the basis of equilibrium constants (K_i values) as activity measurements.[4] Activity cliffs have been systematically identified in publicly available compounds active against current targets.[11,12] Depending on chosen molecular representations, ∼20−35% of all compounds with available high-confidence activity data have been found to participate in the formation of at least one well-defined activity cliff, with MMP-cliffs being the structurally most conservative representation of cliffs.[12]

Following their original definition, activity cliffs have generally been considered at the level of compound pairs, i.e., by separately studying each compound pair forming an "isolated" cliff.[3]

Received: December 9, 2013
Published: January 18, 2014

However, higher-order activity cliff configurations involving multiple highly and lowly potent compounds and "coordinated" activity cliffs have also been detected in compound data sets.[13] In fact, a recent statistical analysis of isolated vs coordinated activity cliffs has revealed that on average more than 95% of all activity cliffs are not formed in isolation but in a coordinated manner.[4] This means that series of compounds with varying potency form multiple overlapping cliffs. Coordinated activity cliffs have higher SAR information content than activity cliffs considered in isolation, which further increases the attractiveness of coordinated cliffs for medicinal chemistry. Hence, the conventional compound pair focus of activity cliff analysis is subject to revision and extension. The prevalence of coordinated activity cliffs implies that many active compounds must participate in the formation of multiple activity cliffs. However, how compounds form such activity cliff configurations and what their sizes and topologies might be is currently unknown. Therefore, we have extracted all activity cliff configurations from currently available bioactive compounds and characterized them in detail. The analysis involved the generation of a global activity cliff network in which cliff configurations form disjoint clusters that can be individually studied.

MATERIALS AND METHODS

Compound Data Sets. Compounds and activity data were assembled from ChEMBL (version 17).[14] We restricted our analysis to compounds with precisely specified equilibrium constants for human targets at the highest confidence level (ChEMBL confidence score 9).[14] A compound with multiple K_i measurements for the same target was only selected if all potency values fell within the same order of magnitude. Then, the average potency was calculated as the final activity annotation. On the basis of these selection criteria, a total of 77 415 compounds were obtained for further analysis. These compounds were active against 661 different targets (with one to 2601 compounds per target set).

MMP-Cliffs. Activity cliffs were defined as MMP-cliffs[10] with a potency difference of at least 2 orders of magnitude between cliff-forming compounds.[4] Transformation size-restricted MMPs[10] were systematically generated for all qualifying compounds using an in-house implementation of the Hussain and Rea algorithm.[9]

Network Analysis. All MMP-cliffs were pooled, and a target-based activity cliff network was generated. In this network, nodes represented cliff-forming compounds and edges, activity cliffs. Network representations were drawn with Cytoscape,[15] and network characteristics[16] were assessed. In addition to global network topology and average node degrees, network "heterogeneity" and "centralization" were calculated as parameters related to the neighborhood of a given node.[17] The network heterogeneity is an index accounting for the variance of connectivity and reflecting the tendency of a network to contain hubs.[17] In addition, the network centralization index is a measure of the centrality of nodes; i.e., it describes the extent to which subsets of nodes are more central than others in the network based on their connectivity. For example, the centralization score of networks with a star-like topology is usually close to 1, whereas the score of uniformly connected networks is close to 0.[17] For activity cliff networks, these indices have been calculated using the Cytoscape NetworkAnalyzer plug-in.[18]

RESULTS AND DISCUSSION

Activity Cliff Statistics. For activity cliff analysis, all bioactive compounds were selected for which high-confidence activity data were available. No data set size restrictions were applied to ensure maximal target coverage. For the 77 415 qualifying compounds, a total of 20 080 MMP-cliffs were identified in 293 target sets. Many small or very small sets did not yield MMPs or MMP-cliffs. These 20 080 activity cliffs included a total of 18 567 unique cliffs detected in one or more target sets. The number of cliffs exclusively identified in a single target set (intraclass cliffs) was 17 287, whereas only 1280 cliffs were found in more than one set (interclass cliffs). Hence, only 6.9% of all MMP-cliffs are multitarget cliffs. The activity cliff statistics are reported in Table 1. The large number of activity cliffs were formed by 11 783 unique compounds, which yielded 14 044 activity cliff compounds. A total of 1766 compounds with multitarget activity participated in the formation of activity cliffs in different target sets (thus rationalizing the difference between the number of unique and cliff-forming compounds). The number of MMP-cliffs per target set varied between one and 1241. Among the cliff-forming compounds, 7358 exclusively acted as highly potent cliff partners, 6414 exclusively as lowly potent partners, and only 272 compounds were found to participate as highly and lowly potent partners in different activity cliffs. All activity cliffs were subjected to network analysis to visualize the configurations they formed.

Table 1. MMP-Cliff Distribution[a]

# MMPs	385 653
# target-based cliffs	20 080 (5.2%)
# unique cliffs	18 567 (4.8%)
# interclass cliffs	1280 (6.9%)
# intraclass cliffs	17 287 (99.7%)

[a]The table reports the total number of MMPs and the number of target-based activity cliffs. In addition, the number of unique activity cliffs (as explained in the text), interclass cliffs (identified in more than one target set), and intraclass cliffs (exclusively identified in a single target set) are reported.

Activity Cliff Network. A global activity cliff network was generated and analyzed in detail. In the network, nodes represented compounds and edges, activity cliffs. The network organized all isolated and coordinated activity cliffs formed by compounds active against 293 targets.

Composition. In Figure 1a, the complete activity cliff network is displayed, which consisted of 14 044 distinct nodes and 20 080 edges. Each edge represented a cliff formed by compounds active against a specific target. For interactive visualization and analysis, the network is also provided as Figure S1 of the Supporting Information. Figure 1a illustrates that the majority of activity cliffs were coordinated, consistent with earlier findings from statistical analysis.[4] Only 769 of all 20 080 activity cliffs were formed in an isolated manner (3.8%), i.e., as compound pairs without structural neighbors forming cliffs. As further discussed below, the degree of coordination among cliffs varied significantly. In the network, coordinated activity cliffs appeared as separate network components of varying size, which we term *activity cliff clusters* in our analysis.

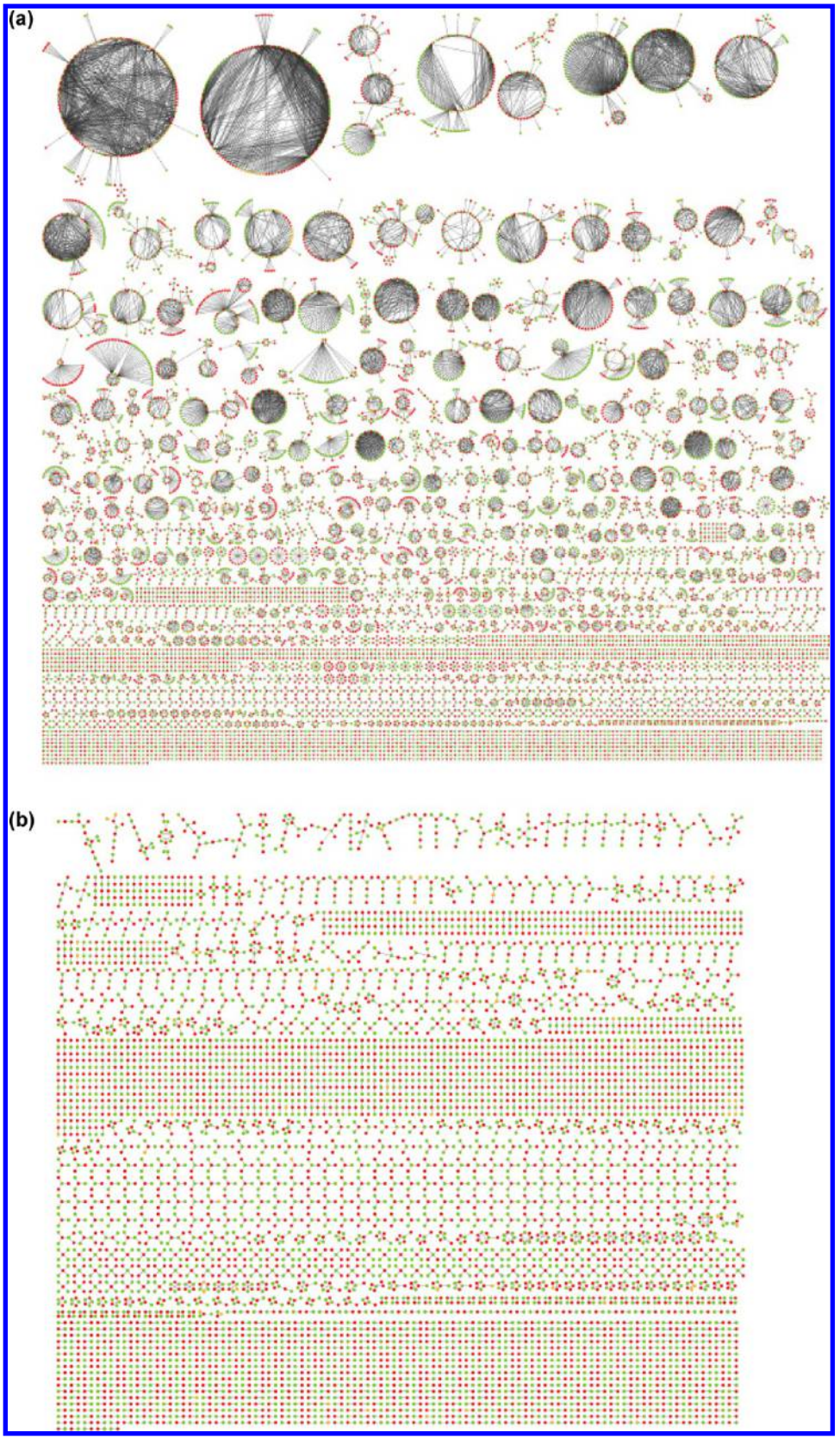


Figure 1. continued

dx.doi.org/10.1021/ci400728r | *J. Chem. Inf. Model.* 2014, 54, 451−461

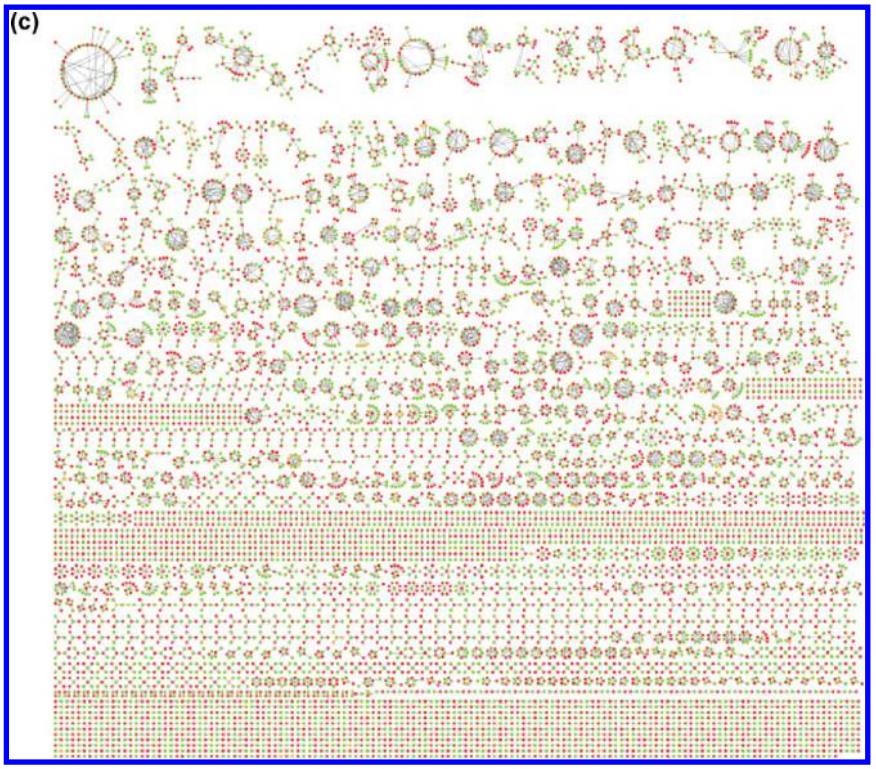


Figure 1. Activity cliff networks. In a, the complete MMP-cliff network is shown. Nodes are colored green if a compound is a highly potent cliff partner, red (lowly potent cliff partner), or yellow if a compound is involved in different activity cliffs both as a highly and lowly potent partner. In b and c, nodes with a degree ≥ 5 and ≥ 10 were removed, respectively, and the network representation was recalculated.

Characterization. The activity cliff network contained a total of 2072 differently sized clusters (including isolated cliffs). The cluster order distribution is reported in Table 2. Cluster order refers to the number of nodes (compounds) per cluster. Twenty-six clusters with more than 50 nodes were detected including four clusters with more than 100 nodes. The largest activity cliff cluster contained 152 nodes. This cluster represented a total of 636 activity cliffs. The overall largest number of cliffs per cluster was 680 detected in another cluster containing 141 nodes. In addition, 420 clusters comprising six to 15 cliff-forming compounds were detected, which also reflected the overall high degree of activity cliff coordination.

Table 2. Activity Cliff Cluster Order Distribution[a]

cluster order	# cluster
1−5	1463
6−10	306
10−15	114
15−20	65
21−30	56
31−40	27
41−50	15
51−60	11
61−70	4
71−80	2
81−90	3
91−100	2
101−152	4

[a]The distribution of activity cliff cluster orders (i.e., numbers of nodes per cluster) across the network is reported.

On the basis of global network analysis, we determined that the union of all clusters followed the power law $P(k) \sim k^{-\gamma}$, with γ having a value of 2.5, which is characteristic of *scale-free* networks that typically yield γ values of 2−3.[16] Here, $P(k)$ is the subset of nodes in the network having k connections to others.

Table 3 reports the node degree distribution in the network. Node degrees varied between 1 and 67, with an average node degree of 2.9. Overall, 1552 nodes with a degree of 5−9 were detected and 496 nodes with a degree of 10−20, revealing the presence of many densely connected nodes. Thus, nodes with a degree ≥ 5 were considered *activity cliff hubs*. In total, the network contained 2166 (15.4%) and 614 (4.4%) hubs with a degree ≥ 5 and a degree ≥ 10, respectively.

Table 3. Node Degree Distribution[a]

node degree	# nodes
1−4	11878
5−9	1552
10−14	341
15−20	155
21−30	85
31−40	17
41−50	9
51−60	4
61−70	3

[a]The node degree distribution of the activity cliff network is reported.

Modification. In the activity cliff network, there were 463 (22.3%) clusters containing at least one hub including 116 clusters with at least one hub with a degree ≥ 10. Thus, activity cliff hubs were integral components of the network. To evaluate the role of these hubs for the network and its global topology, two

dx.doi.org/10.1021/ci400728r | J. Chem. Inf. Model. 2014, 54, 451−461

network variants were generated after removal of hubs with a degree ≥ 5 and ≥ 10, respectively. These network variants are displayed in Figure 1b and c. In addition, a comparison of the statistics for the original network and its two variants is presented in Table 4. The absence of increasing numbers of hubs led to increasing randomness of the network variants. As expected, these modifications reduced network heterogeneity, a measure for the tendency of a network to contain hubs, and also network centralization, thus indicating increasingly uniform connectivity. Taken together, these findings were also consistent with the global scale-free character of the original activity cliff network.

Table 4. Activity Cliff Network Statistics[a]

network statistics	complete network	subnetwork 1: no hubs with degree ≥ 5	subnetwork 2: no hubs with degree ≥ 10
# clusters	2072	2171	2173
# nodes	14 044	7265 (51.7%)	11 115 (79.1%)
# edges	20 080	5508 (27.4%)	11 381 (56.7%)
average node degree	2.9	1.5	2.0
network heterogeneity	1.3	0.5	0.79
network centralization	0.005	0	0.001

[a]Subnetworks 1 and 2 were derived from the original activity cliff network by removal of nodes with a degree of five and 10 or more, respectively, followed by recalculation of the network representation.

Activity Cliff Cluster Topology. Next, we systematically determined activity cliff cluster topologies present in the network and compared topological features. Given the very low proportion of multitarget (interclass) activity cliffs formed by ChEMBL compounds, as reported above, all recurrent topologies identified in our analysis are formed by single-target (intraclass) activity cliffs.

Distribution. Table 5 reports the cluster topology distribution of the activity cliff network. The global network consisted of 2072 activity cliff clusters that represented 450 distinct cluster topologies. A small set of 39 cluster topologies (including isolated cliffs) accounted for 1630 different clusters. These clusters were of small to moderate size (with up to 12 compounds) and contained a total of 5323 compounds forming 3866 activity cliffs. In addition, 416 clusters with fewer than 50 compounds yielded 385 distinct topologies. Hence, these topologies were detected only once or twice. The corresponding clusters contained 6722 compounds that formed ~55% (11 083) of all activity cliffs. Furthermore, there were 26 large clusters with unique topologies that contained a total of 1999 compounds forming 5131 cliffs. As further discussed below, large clusters were typically characterized by a high degree of activity cliff density.

Table 5. Cluster Topology Distribution[a]

topology	# clusters	# compounds	# cliffs
39 topologies instances ≥ 3	1630	5323	3866
26 topologies instances = 1 # compounds > 50	26	1999	5131
385 topologies instances < 3 # compounds ≤ 50	416	6722	11 083
total			
450 topologies	2072	14 044	20 080

[a]All activity cliff clusters are organized according to their topologies and compound composition.

The cluster frequency distribution of unique topologies is reported in Table 6. A total of 381 clusters with unique topology were observed only once, whereas six and seven distinct topologies were each observed 10 or more and 20 or more times, respectively (i.e., in the latter case, 20 or more clusters were found to share one of seven unique topologies). Thirteen of the 39 topologies that were observed at least three times contained a hub with a degree ≥ 5.

Table 6. Cluster Frequency for Unique Topologies[a]

cluster frequency	# topologies
≥ 20	7
≥ 10	6
≥ 5	10
≥ 3	16
= 2	30
= 1	381

[a]The cluster frequency for unique topologies is reported. For example, the first row of the table means that seven distinct topologies were each represented by at least 20 different clusters.

Topological Categories. We determined that the 1630 clusters with recurrent topology could be assigned to only three main topology categories and a limited number of extensions of these categories, as schematically shown in Figure 2. The main categories included the *star*, *chain*, and *rectangle* topologies. For the *star* and *chain* topologies, frequently observed extensions included the *twin star* and *modified chain*, respectively. In addition, the *rectangle* topology had three well-defined extensions including the *modified*, *nested*, and *fused rectangle*, as illustrated in Figure 2. Taken together, this limited set of topologies or combinations of these topologies covered all small and moderately sized activity cliff clusters for compounds active against current targets.

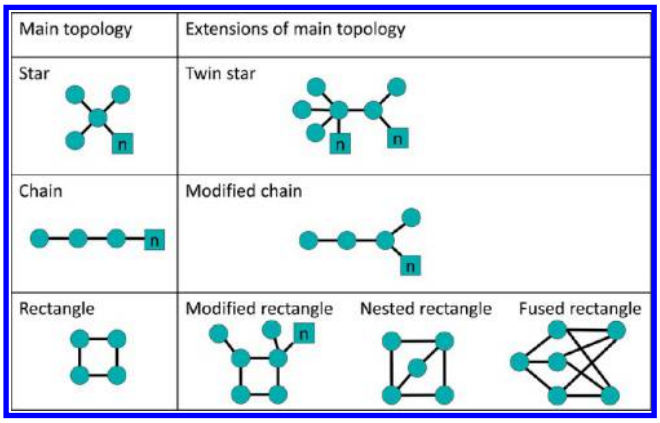


Figure 2. Topology categories. The three most frequently observed activity cliff cluster topologies (left) and their extensions (right) are schematically illustrated. The three main topology categories were termed *star*, *chain*, and *rectangle*, respectively. Squared nodes represent variable node numbers (*n*).

Complex Topologies. Figure 3 shows examples from the set of the 26 largest clusters with unique topology. The cluster in Figure 3a consists of 56 opioid receptor agonists forming 70 activity cliffs. This cluster combines different star and rectangle

dx.doi.org/10.1021/ci400728r | J. Chem. Inf. Model. 2014, 54, 451−461

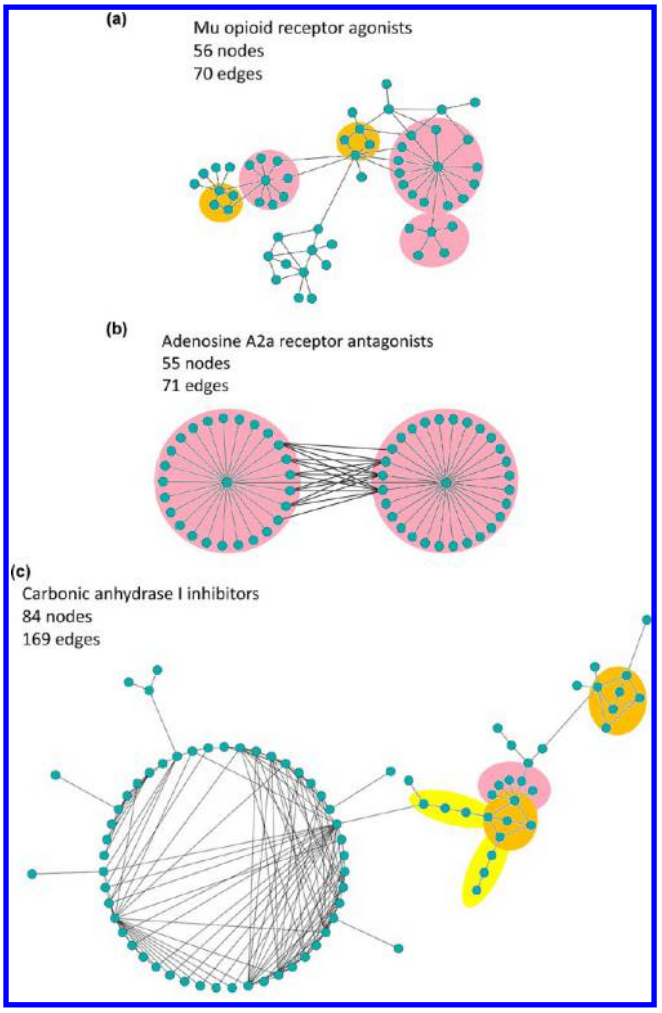


Figure 3. Large activity cliff clusters. Three exemplary large activity clusters with unique topology are shown in detail. These clusters comprise 56 (a), 55 (b), and 84 (c) compounds with different specific activities that form 70, 71, and 169 activity cliffs, respectively. Recurrent topological motifs are highlighted in yellow (*chains*), orange (*rectangles*), and pink (*stars*).

motifs with another less well-defined motif. Combinations of the three main topologies were often observed in large clusters. Similarly, the cluster consisting of 55 adenosine receptor antagonists forming 71 cliffs in Figure 3b displays a further extended twin star topology with 14 edges connecting the two star motifs. Furthermore, the cluster in Figure 3c (with 84 carbonic anhydrase inhibitors forming 169 cliffs) also contains a peripheral combination of star, rectangle, and chain motifs. In addition, its central component is characterized by a high density of activity cliffs, giving rise to a complex topology that is difficult to resolve into individual motifs. Such central components with a high density of activity cliffs were characteristics of the largest clusters in the network. The circle layout of these densely connected components was generated to accommodate high activity cliff density and hence does not represent an independent topology. The target distribution of activity cliff clusters and topologies is reported below.

Frequency of Occurrence. Table 7 reports the most frequently observed topologies (including isolated cliffs) and cluster size variations. With 335 instances, chains with three compounds represented the most frequent activity cliff clusters, followed by stars with four (122 instances) and five compounds (70), chains with four (53), and twin stars with five compounds (43). The basic rectangle consisting of four compounds was 26 times observed. Thus, the most frequently occurring activity cliff clusters were of relatively small size. Larger clusters that were also at least three times observed included, among others, stars with 12, twin stars with 11, or modified rectangles with nine compounds.

As also reported in Table 7, chains and stars were much more frequently observed than rectangles. Overall, there were 243 instances of stars covering a total of 1222 compounds and 979 cliffs and 388 instances of chains with 1217 compounds and 829 cliffs. Hence, chains with three or four nodes were the overall most frequent topologies. However, clusters with star and chain topologies had very similar compound coverage. Stars

Table 7. Frequently Occurring Topologies of Activity Cliff Clusters of Varying Size[a]

# instances	topology category	# cpds per topology	# target sets
769	*isolated cliff*	2	202
335	*chain*	3	127
122	*star*	4	74
70	*star*	5	53
53	*chain*	4	42
43	*twin star*	5	39
26	*rectangle*	4	22
19	*star*	6	17
18	*mod. rect.*	5	14
16	*nested rect.*	5	15
14	*twin star*	6	14
11	*star*	8	11
10	*star*	7	9
9	*twin star*	7	9
8	*nested rect.*	6	8
8	*nested rect.*	6	7
8	*twin star*	7	8
7	*chain*	5	7
7	*twin star*	6	7
6	*mod. rect.*	6	6
6	*twin star*	8	6
5	*mod. chain*	6	5
5	*twin star*	7	5
5	*fused rect.*	6	5
4	*mod. rect.*	6	4
4	*mod. rect.*	7	4
4	*twin star*	8	4
4	*star*	9	4
4	*star*	10	4
3	*twin star*	6	3
3	*twin star*	7	3
3	*mod. rect.*	7	3
3	*nested rect.*	7	3
3	*mod. rect.*	7	2
3	*mod. rect.*	8	3
3	*twin star*	9	3
3	*mod. rect.*	9	3
3	*twin star*	11	3
3	*star*	12	3

[a]The first row reports that a total of 769 isolated activity cliffs, which consisted of two compounds (cpds), were found in 202 target sets. In the second row, it is reported that an activity cliff "chain" containing nine compounds was 335 times detected across 127 target sets. "mod." stands for modified and "rect." for rectangle.

dx.doi.org/10.1021/ci400728r | *J. Chem. Inf. Model.* 2014, 54, 451−461

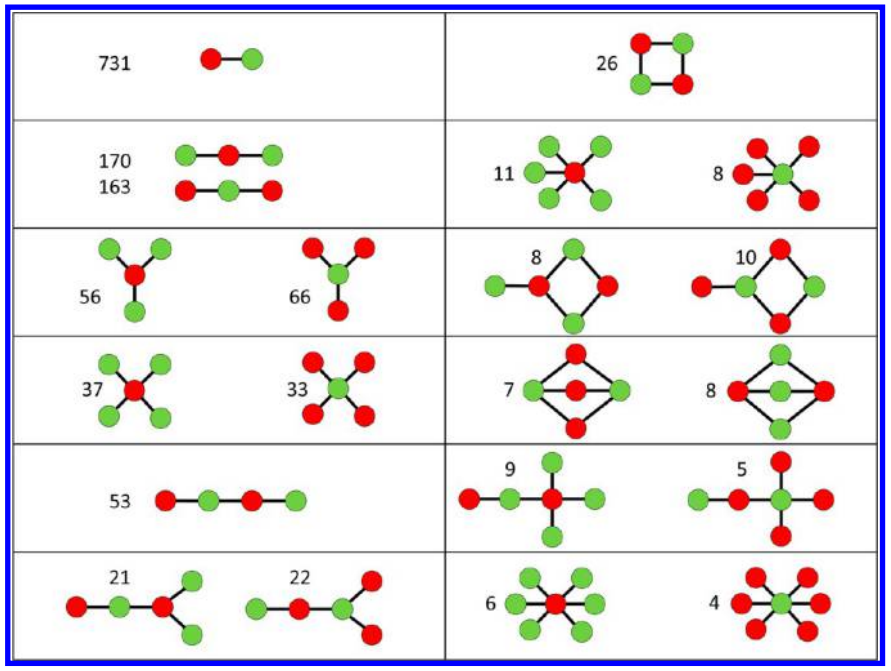


Figure 4. Frequently occurring activity cliff cluster topologies. Shown are the 12 most frequently observed cluster topologies. The number of occurrences is reported. Nodes representing highly and lowly potent cliff partners are colored red and green, respectively. Compounds that were both highly and lowly potent partners in different activity cliffs did not occur in clusters having the most frequent topologies (all of which represented topologies and extensions depicted in Figure 2).

Table 8. Target Sets with Largest Numbers of Isolated Activity Cliffs[a]

# isolated cliffs	ChEMBL TID	target name	target family	# active compounds	# cliff compounds
24	234	dopamine D3 receptor	monoamine receptor GPCR family 1	1384	259
22	233	mu opioid receptor	short peptide GPCR family 1	1582	359
21	3594	carbonic anhydrase IX	carbonic anhydrase family	1313	143
20	217	dopamine D2 receptor	monoamine receptor GPCR family 1	2038	227
20	261	carbonic anhydrase I	carbonic anhydrase family	1656	360
19	226	adenosine A1 receptor	nucleotide-like receptor GPCR family 1	2172	283
19	264	histamine H3 receptor	monoamine receptor GPCR family 1	2012	319
18	205	carbonic anhydrase II	carbonic anhydrase family	1697	276
18	237	kappa opioid receptor	short peptide GPCR family 1	1491	451
17	253	cannabinoid CB2 receptor	lipid-like ligand receptor GPCR family 1	1994	504

[a]Target-based compound data sets, their number of activity cliff-forming compounds, and the number of isolated activity cliffs they contain are reported. TID stands for target ID.

represented activity cliff configurations resulting from combinations of a highly potent compound with multiple lowly potent analogs and vice versa. In medicinal chemistry, such compound series are likely to originate from hit-to-lead and lead optimization efforts. Clusters with star topology included many hubs and were found to be mostly responsible for the global scale-free character of the activity cliff network.

Figure 4 shows the 12 most frequently occurring cluster topologies (including isolated cliffs). For nine of the 11 multicliff topologies, different subsets with alternative arrangements of highly and lowly potent activity cliff partners were identified. The exceptions were isolated cliffs, the rectangle, and chain with four compounds each for which no alternative subsets were possible. Small chains with three compounds representing two activity cliffs formed by two highly and one lowly potent cliff partner (170 instances) or one highly potent and two lowly potent partners (163) dominated the topology distribution. In addition, stars with four compounds representing three cliffs or five compounds representing four cliffs were also frequently observed, with a total of 122 and 70 instances, respectively. These topologies required the combination of a highly potent cliff compound with three or four lowly potent ones or vice versa. Chains with four compounds including two highly and two lowly potent cliff partners that formed three activity cliffs were detected 53 times (compared to 26 instances of the basic rectangle having the same compound composition but forming four cliffs). As shown in Figure 4, topology extensions such as modified and nested rectangles or modified chains were also recurrent.

Target Distribution. As reported in Table 7, cluster topologies were widely distributed over different target sets. Stars, chains, and rectangles were found in 142, 144, and 70 target sets, respectively. Isolated activity cliffs were detected in 202 target sets, but their frequency of occurrence was very low. Only 21 of these 202 sets contained more than 10 isolated cliffs. Table 8 lists the top 10 target sets with most isolated activity cliffs. The maximum number of isolated cliffs per set was 24. This target set consisted of 1384 dopamine D3 receptor antagonists, 259 of which participated in the formation of activity cliffs.

dx.doi.org/10.1021/ci400728r | J. Chem. Inf. Model. 2014, 54, 451−461

Table 9. Target Sets with Largest Numbers of Different Activity Cliff Cluster Topologies[a]

# cluster topologies	ChEMBL TID	target name	target family	# active compounds	# cliff compounds
33	251	adenosine A2a receptor	nucleotide-like receptor GPCR family 1	2601	496
33	253	cannabinoid CB2 receptor	lipid-like ligand receptor GPCR family 1	1994	504
25	218	cannabinoid CB1 receptor	lipid-like ligand receptor GPCR family 1	1760	342
23	256	adenosine A3 receptor	nucleotide-like receptor GPCR family 1	2095	566
22	226	adenosine A1 receptor	nucleotide-like receptor GPCR family 1	2172	283
22	264	histamine H3 receptor	monoamine receptor GPCR family 1	2012	319
21	217	dopamine D2 receptor	monoamine receptor GPCR family 1	2038	227
20	233	mu opioid receptor	short peptide GPCR family 1	1582	359
18	234	dopamine D3 receptor	monoamine receptor GPCR family 1	1384	259
18	236	delta opioid receptor	short peptide GPCR family 1	1315	238

[a]Target-based compound data sets, their number of activity cliff-forming compounds, and different activity cluster topologies are reported. TID stands for target ID.

Table 10. Clusters with Largest Numbers of Activity Cliffs[a]

# cliffs	ChEMBL TID	target name	target family	# cluster compounds
680	237	kappa opioid receptor	short peptide GPCR family 1	141
636	256	adenosine A3 receptor	nucleotide-like receptor GPCR family 1	152
364	244	coagulation factor X	serine protease family	74
330	244	coagulation factor X	serine protease family	88
292	255	adenosine A2b receptor	nucleotide-like receptor GPCR family 1	88
283	256	adenosine A3 receptor	nucleotide-like receptor GPCR family 1	133
224	2014	nociceptin receptor	short peptide GPCR family 1	78
211	244	coagulation factor X	serine protease family	104
180	259	melanocortin receptor 4	short peptide GPCR family 1	53
169	261	carbonic anhydrase I	carbonic anhydrase family	84
168	259	melanocortin receptor 4	short peptide GPCR family 1	96
148	4409	phosphodiesterase 10A	phosphodiesterase family	54
147	204	thrombin	serine protease family	69
141	1914	butyrylcholinesterase	type-B carboxylesterase/lipase family	65
127	205	carbonic anhydrase II	carbonic anhydrase family	91
126	222	norepinephrine transporter	sodium:neurotransmitter symporter (SNF) family	53
117	1855	gonadotropin-releasing hormone receptor	short peptide GPCR family 1	68
108	249	neurokinin 1 receptor	short peptide GPCR family 1	53
107	259	melanocortin receptor 4	short peptide GPCR family 1	52
105	1997	equilibrative nucleoside transporter 1	SLC29A/ENT transporter family	70

[a]The top 20 clusters with largest numbers of activity cliffs and their composition are reported. TID stands for target ID.

Many target sets yielded activity cliff clusters with different topologies. In 30 target sets, at least 10 different topologies were detected. Table 9 lists the top 10 target sets having the largest number of cluster topologies. Both adenosine A2 and cannabinoid CB2 receptor antagonists yielded 33 different cluster topologies. These data sets consisted of 2601 and 1994 compounds, 496 and 504 of which formed activity cliffs, respectively. Target sets containing at least 100 compounds yielded seven to 13 different cluster topologies. In addition, sets containing at least 200 compounds yielded 10 to 25 topologies and sets with more than 400 compounds, 16 to 33 topologies. Thus, as would be expected, the number of cluster topologies generally increased with the size of data sets and the number of activity cliff-forming compounds.

Table 10 reports clusters with the largest numbers of activity cliffs and their target distribution. As discussed above, the 26 largest clusters alone contained 1999 cliff-forming compounds (14.2%) and accounted for 5131 activity cliffs (25.6% of all cliffs). Hence, these clusters were centers of coordinated activity cliff formation in the global network. The largest clusters originated from a small number of target sets. For example, the clusters in Table 10 included three different clusters of coagulation factor Xa inhibitors and three others with different adenosine receptor ligands. Overall, ligands of different G protein coupled receptors formed the majority of large activity cliff clusters.

Interpretation and Utility. Activity cliff cluster topologies were extracted from network representations. At the level of subgraphs, cluster topologies can be systematically compared. As we have shown, a comparison at this level is sufficient to obtain a detailed view of currently available cluster topologies, which were thus far unknown. However, one can go beyond this level and view the compounds forming clusters of interesting topology. From compounds forming coordinated activity cliffs within a cluster, SAR information can be directly obtained and substituents can be identified that are responsible for activity cliff formation or characteristic of highly potent compounds. Thus, cluster topologies provide immediate access to SAR exploration, and characteristic topologies such as stars or rectangles can be prioritized depending on the specific applications. Two representative examples for SAR analysis from activity cliff clusters of defined topology are provided in Figure 5. Furthermore, for the analysis of coordinated activity cliffs, which is much more complex than cliff assessment at the level of compound pairs, network-derived topologies have the

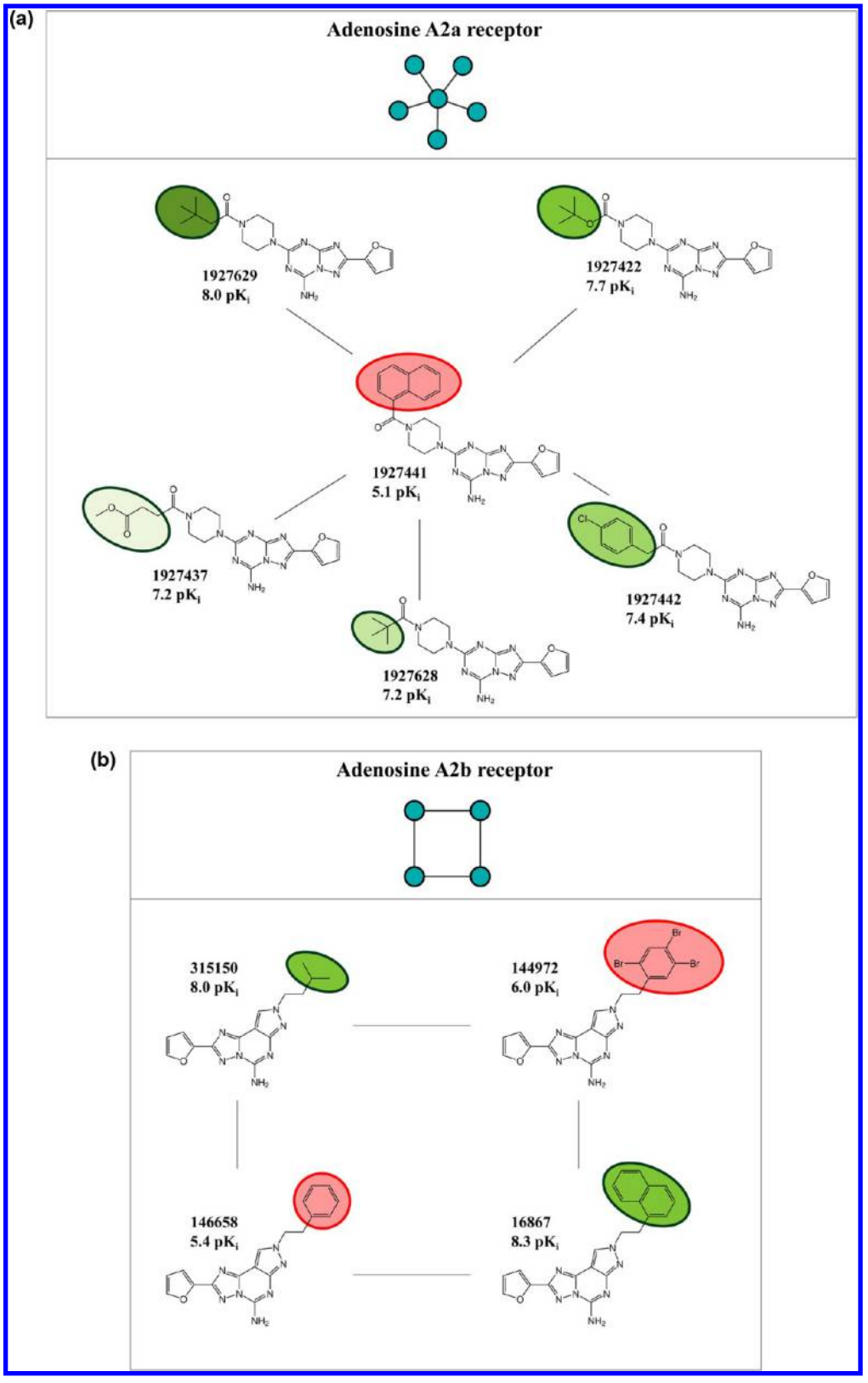


Figure 5. Exemplary cluster topologies and compounds. Shown are compounds forming representative activity cliffs of exemplary cluster topologies (a star, b rectangle). Structural modifications are highlighted and colored according to compound activity (red, low activity; green, high activity). Examples are taken from the (a) adenosine A2a receptor and (b) adenosine A2b receptor target sets.

advantage that they immediately reveal all compounds participating in a cluster and all activity cliffs comprising the cluster, as illustrated in Figure 5, without the need to search for additional compounds that might further extend given activity cliff(s). Moreover, a cluster of similar size sharing the same topology can be selected from a given target set and analyzed in a comparative manner, which further increases the amount of SAR information that can be extracted in an organized manner from

dx.doi.org/10.1021/ci400728r | J. Chem. Inf. Model. 2014, 54, 451−461

structurally heterogeneous sets of specifically active compounds. SAR information obtained from the topology-supported analysis of coordinated activity cliffs might then be utilized in the context of compound exploration and optimization.

CONCLUSIONS

Following their original definition, activity cliffs have predominantly been studied at the level of individual compound pairs. As has recently been shown, most activity cliffs are not formed in isolation but as higher-order configurations involving multiple cliffs. However, the composition, structure, and topology of these activity cliff arrangements are currently unknown. Our analysis provides a first view of activity cliff cluster topologies. The analysis is descriptive in nature, aiming at providing a comprehensive account of activity cluster topologies found in currently available bioactive compounds. To elucidate activity cliff configurations, cliffs have been systematically extracted from bioactive compounds on the basis of high-confidence activity data. Maximal target coverage of compound data sets was ensured. More than 20 000 well-defined activity cliffs were obtained that were formed by compounds active against nearly 300 different targets (to our knowledge, the largest collection of activity cliffs studied to date). A global target-based activity cliff network was generated to identify and visualize all cliff configurations. In the network, activity cliff configurations formed individual clusters that were systematically analyzed. The network provides a basis of the extraction and further characterization of activity cliff cluster topologies. Activity cliff clusters of very different sizes were identified that were widely distributed over target sets. A limited number of very large clusters with complex topology was formed that represented centers of coordinated activity cliff formation and originated from a small number of target sets. However, most clusters were of small to moderate size and characterized by only three basic topologies and several extensions of these topologies. Large activity cliff clusters often contained different combinations of these basic topological motifs. Small clusters with chain topology were overall most frequently observed. These clusters were produced by series of pairwise analogs with significantly varying potency. Clusters with star topology were largely responsible for the scale-free nature of the global activity cliff network that contained many cliff-forming hubs. Star topology of clusters resulted from the presence of a highly potent compound forming multiple activity cliffs with lowly potent partners and vice versa. A characteristic feature of activity cliff clusters with frequently observed topology was that they did not contain compounds involved in multiple activity cliffs as both highly and lowly potent cliff partners. Thus, compounds with variable potency relationships were rare within activity cliff clusters. In general, activity cliff clusters have higher SAR information content than isolated cliffs and are thus of particular interest for large-scale SAR exploration. The finding that small to moderately sized activity cliff clusters had well-defined topologies across many different target sets implied that structure−activity relationships captured by these types of clusters might often be similar. This represents an interesting aspect for the study of activity cliff configurations from a medicinal chemistry perspective. Taken together, the results of our analysis have provided a detailed view of activity cliff configurations formed by compounds active against the current spectrum of targets.

ASSOCIATED CONTENT

Supporting Information

Supporting Figure S1 shows original and modified activity cliff networks. This information is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

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

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Ye Hu for help with network calculations. D.S. is supported by *Sonderforschungsbereich 704* of the *Deutsche Forschungsgemeinschaft*.

REFERENCES

(1) Maggiora, G. M. On Outliers and Activity Cliffs − Why QSAR often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535−1535.
(2) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J. Med. Chem.* **2012**, *55*, 2932−2942.
(3) Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. *J. Med. Chem.* In press, DOI: 10.1021/jm401120g.
(4) Hu, Y.; Stumpfe, D.; Bajorath, J. Advancing the Activity Cliff Concept [v1; ref. status: indexed, http://f1000r.es/1wf]. *F1000Research* **2013**, *2*, 199 (DOI: 10.12688/f1000research.2-199.v1).
(5) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity Landscape Representations for Structure-Activity Relationship Analysis. *J. Med. Chem.* **2010**, *53*, 8209−8223.
(6) Peltason, L.; Iyer, P.; Bajorath, J. Rationalizing Three-Dimensional Activity Landscapes and the Influence of Molecular Representations on Landscape Topology and the Formation of Activity Cliffs. *J. Chem. Inf. Model.* **2010**, *50*, 1021−1033.
(7) 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.
(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) 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.
(12) Stumpfe, D.; Bajorath, J. Frequency of Occurrence and Potency Range Distribution of Activity Cliffs in Bioactive Compounds. *J. Chem. Inf. Model.* **2012**, *52*, 2348−2353.
(13) 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.
(14) 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.
(15) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A

Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13*, 2498−2504.
(16) Barabási, A. L.; Oltvai, Z. N. Network Biology: Understanding the Cell's Functional Organization. *Nat. Rev. Genet.* **2004**, *5*, 101−113.
(17) Dong, J.; Horvath, S. Understanding Network Concepts in Modules. *BMC Syst. Biol.* **2007**, *1*, 24.
(18) Doncheva, N. T.; Assenov, Y.; Domingues, F. S.; Albrecht, M. Topological Analysis and Interactive Visualization of Biological Networks and Protein Structures. *Nat. Protoc.* **2012**, *7*, 670−685.