JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

MMP-Cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs

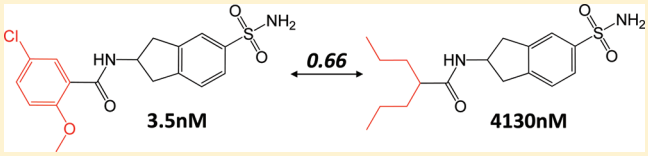
Xiaoying Hu,[†,‡] Ye Hu,[†] Martin Vogt,[†] Dagmar Stumpfe,[†] and Jürgen Bajorath[†,*]

[†]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.

[‡]State Key Laboratory of Chemical Resource Engineering, Department of Pharmaceutical Engineering, P.O. Box 53, Beijing University of Chemical Technology, 15 BeiSanHuan East Road, Beijing 100029, People's Republic of China

Supporting Information

ABSTRACT: Activity cliffs are generally defined as pairs of structurally similar compounds having large differences in potency. The analysis of activity cliffs is of general interest because structure−activity relationship (SAR) determinants can often be deduced from them. Critical questions for the study of activity cliffs include how similar compounds should be to qualify as cliff partners, how similarity should be assessed, and how large potency differences between participating compounds should be. Thus far, activity cliffs have mostly been defined on the basis of calculated Tanimoto similarity values using structural descriptors, especially 2D fingerprints. As any theoretical assessment of molecular similarity, this approach has its limitations. For example, calculated Tanimoto similarities might often be difficult to reconcile and interpret from a chemical perspective, a point of critique frequently raised in medicinal chemistry. Herein, we have explored activity cliffs by considering well-defined substructure replacements instead of calculated similarity values. For this purpose, the matched molecular pair (MMP) formalism has been applied. MMPs were systematically derived from public domain compounds, and activity cliffs were extracted from them, termed MMP-cliffs. The frequency of cliff formation was determined for compounds active against different targets, MMP-cliffs were analyzed in detail, and re-evaluated on the basis of Tanimoto similarity. In many instances, chemically intuitive activity cliffs were only detected on the basis of MMPs, but not Tanimoto similarity.



INTRODUCTION

Activity cliffs are the most prominent features of activity landscapes[1−3] of compound data sets.[3−5] They are often thought to be a particularly rich source of SAR information, given that small chemical changes elicit large biological effects. However, the study of activity cliffs alone is not sufficient for comprehensive SAR analysis[2] because of the often complex nature of SAR patterns in compound data sets.[2,3] Nevertheless, activity cliffs usually are the primary focal points of activity landscape exploration[3,4] and of high interest for medicinal chemistry.

Despite its intuitive nature, the activity cliff concept has its caveats. Activity cliffs can be defined in different ways, either as discrete states[2] or as a continuum of similarity/potency relationships.[2,4] As long as activity cliffs are considered as discrete states, similarity and potency criteria for cliff formation must be clearly specified and consistently applied.[2] Furthermore, activity cliff distributions are often significantly affected by the type and quality of experimental measurements that are available.[6] In many instances, activity cliffs are not conserved when alternative potency measurements are used.[6] Moreover, chosen molecular representations and similarity metrics strongly influence the assessment of activity cliffs,[3,7] more so than experimental data variability. Activity cliffs have thus far mostly been studied on the basis of Tanimoto similarity[8] calculations using various molecular descriptors. Depending on the chosen molecular representation, compounds might be classified as more or less similar and, consequently, activity cliffs identified in a given descriptor reference space might not be conserved in another.[7] In addition to the representation/metric dependence, the similarity caveat has yet another dimension. Calculated similarity values, a hallmark of many chemoinformatics applications, are often questioned in medicinal chemistry because of their limited chemical interpretability. Whether or not compound pairs yielding different similarity values are more or less similar to each other is often rather difficult to rationalize from a chemical perspective. These potential ambiguities also affect the analysis of activity cliffs.[2,3]

In order to address the similarity conundrum and generalize the description of activity cliffs, it would be highly desirable to consider robust similarity measures that reduce the representation dependence of cliff formation as much as possible and support interpretability. This might be accomplished, for example, by replacing calculated similarity values with defined substructure relationships, as has been attempted to improve the interpretability of SAR networks.[9]

Therefore, we have explored an alternative way to define activity cliffs by using substructure relationships as a similarity criterion. To generalize the approach, we have utilized the

Received: March 1, 2012
Published: April 10, 2012

Table 1. Compound Data, MMPs, and Activity Cliffs

Data source	# Sets	# Cpds	# MMPs	# MCs	# Cpds forming MCs	# Sets (≥1 MC)	# Sets (≥5 MCs)
BindingDB	621	59,050	314,209	21,285	13,980	361	241

The numbers of target sets (# Sets), compounds (# Cpds), and resulting unique MMPs are reported. Furthermore, the numbers of MMP-cliffs (# MCs) and compounds forming these cliffs are provided. In addition, the numbers of target sets containing at least one (≥1 MC) and five (≥5 MCs) MMP-cliffs are reported.

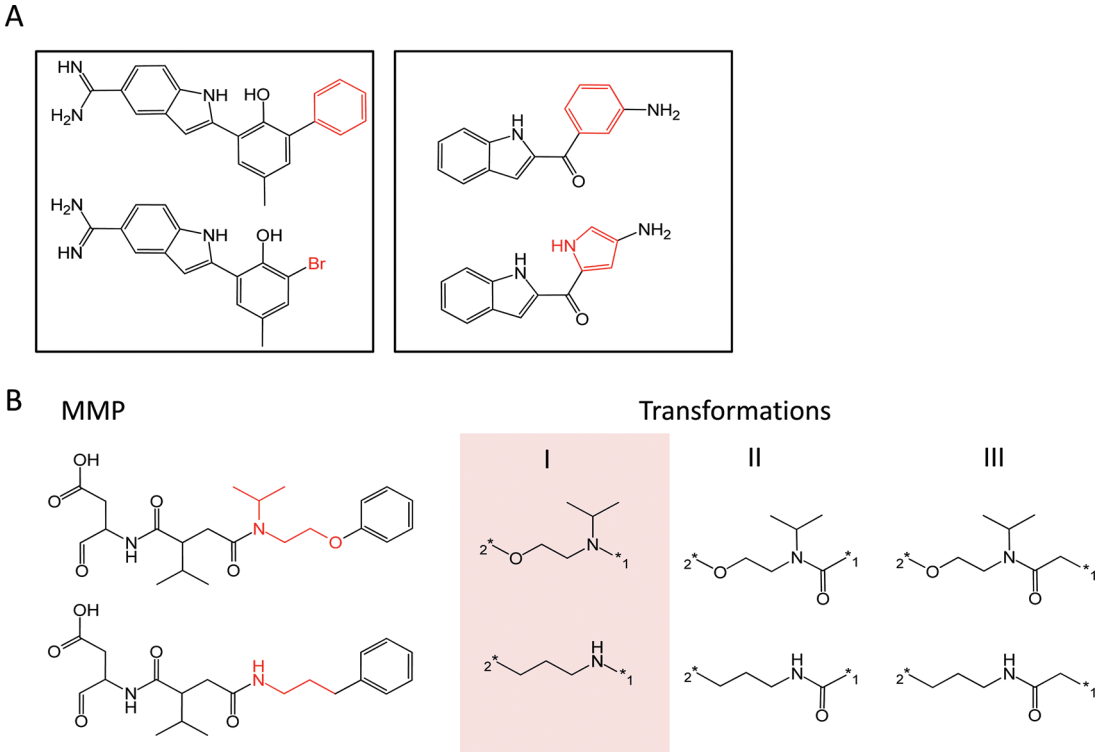


Figure 1. Exemplary MMPs. (A) Shown are two representative compound pairs that form MMPs. Exchanged fragments are colored red. (B) For the MMP on the left, all possible transformations of different sizes are shown on the right. The smallest transformation selected to define the MMP is highlighted.

MMP formalism.[10] An MMP is defined as a pair of compounds that only differ at a single site (represented by a substructure) such as a ring or an R-group. Previously, we have applied the MMP concept to search for small chemical substitutions that display a general tendency to form activity cliffs across different target families.[11] Herein, we have determined activity cliffs in public domain compounds by systematically exploring substructure transformations, leading to the introduction of MMP-cliffs that are often undetected when similarity relationships are numerically quantified.

MATERIALS AND METHODS

Compound Data. For our analysis, the two major public domain repositories of compound optimization data from medicinal chemistry sources were considered, ChEMBL[12,13] and BindingDB.[14,15] PubChem BioAssay Data[16] were previously found to be a negligible source of activity cliffs spanning at least 2 orders of magnitude in potency.[17] We determined that the current (February 2012) version of BindingDB contains 80.3% of the ChEMBL compounds and hence focused our analysis on BindingDB and its target classification. Taking the likely influence of measurement variability on activity cliff formation into account[6] and to ensure a high level of data integrity, only compounds were selected for further analysis for

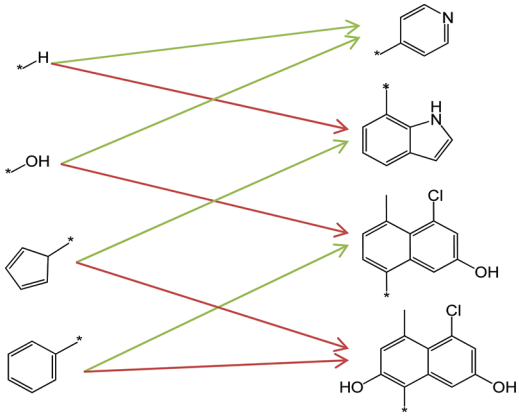


Figure 2. Cliff transformations of maximal size. Shown are eight hypothetical chemical transformations. Given the size restrictions defined for activity cliff forming transformations, only transformations indicated by green arrows qualify as largest permitted chemical changes. By contrast, transformations indicated by red arrows are not permitted.

which explicit K_i measurements were available and which had at least 10 μM potency. If multiple K_i values were available, their geometric mean was calculated to yield the final potency

dx.doi.org/10.1021/ci3001138 | J. Chem. Inf. Model. 2012, 52, 1138−1145

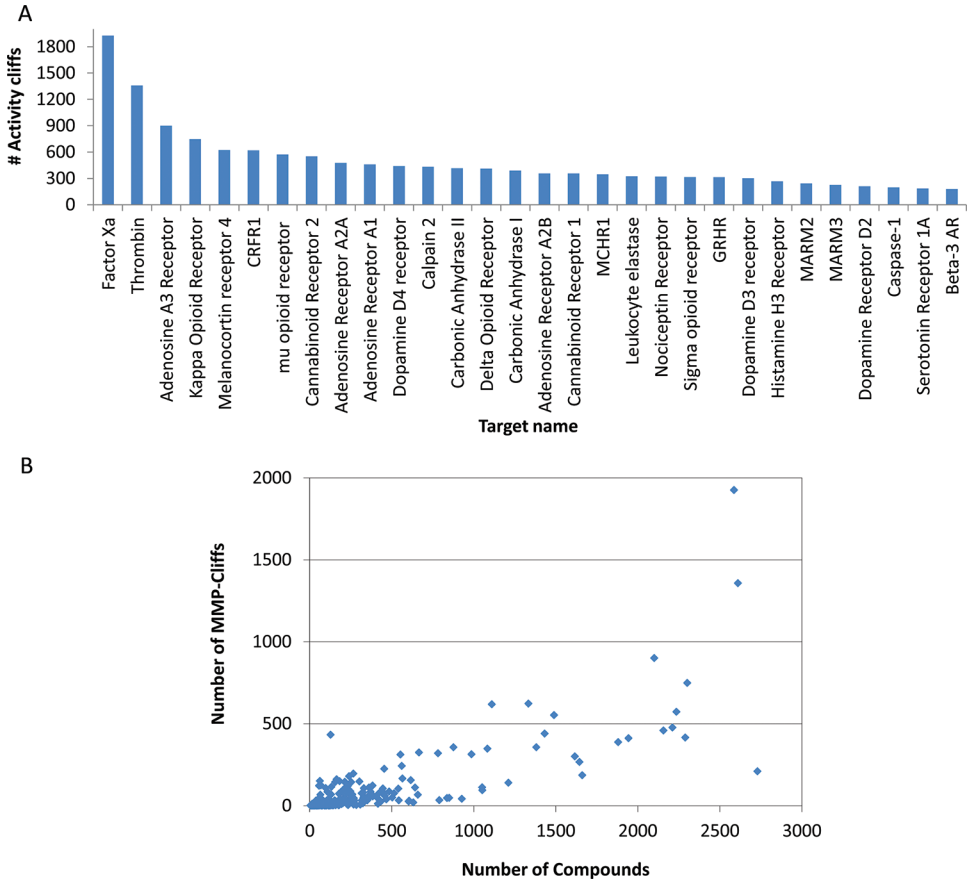


Figure 3. Distribution of MMP-cliffs. (A) The top 30 target sets containing the largest numbers of MMP-cliffs are reported. Target abbreviations: CRFR1, corticotropin releasing factor receptor 1; MCHR1, melanin-concentrating hormone receptor 1; GRHR, gonadotropin-releasing hormone receptor; MARM2, muscarinic acetylcholine receptor M2; MARM3, muscarinic acetylcholine receptor M3; and Beta-3 AR, beta-3 adrenergic receptor. (B) The number of MMP-cliffs formed in each target set is reported as a function of the number of compounds per set.

annotation of a compound. The use of equilibrium constants generally produces most reliable activity cliff assignments.[6] Accordingly, a total of 59,050 compounds were selected from BindingDB that were active against 621 human targets (i.e., forming 621 different target sets containing at least five compounds) (Table 1).

MMP Generation. The MMP concept is illustrated in Figure 1. Compounds forming an MMP are related by the exchange of a pair of distinguishing substructures that might include terminal groups or central fragments of varying size. MMP generation was facilitated using an in-house implementation of the algorithm by Hussain and Rea.[18] All single non-ring bonds between two non-hydrogen atoms in a molecule were subjected to systematic deletion of individual bonds as well as two- and three-bond combinations, which resulted in different numbers of fragments. If a single bond was deleted, the compound was separated into two fragments. Each of these fragments was inserted once as a key in an index table and the other as its value. If two molecules forming an MMP differed only in a single group attached to a common core via a single bond, the two groups were associated with the same key (common core). Accordingly, all MMPs can be identified in the index table by searching for keys with more than one value. In addition to single bonds, bond pairs and triplets were also deleted, which resulted in a core fragment and two or three terminal substituents. These groups of substituents were then stored as a key and the corresponding core as the value. For all fragments, connectivity information was recorded, and the index table was processed by identifying all pairs of compounds that shared a key and had different value(s), thus meeting the condition for MMP formation. The underlying substructure exchange is often termed a (chemical) transformation. In addition, we restricted MMP formation by the requirement that keys must have at least twice the size of value fragments. Furthermore, if several transformations yielded the same MMP, only the smallest transformation was retained, as also illustrated in Figure 1. This was done to minimize the size of structural modifications for activity cliff consideration. If not stated otherwise, all specific atom counts reported herein refer to non-hydrogen atoms.

Activity Cliff Definition. For the formation of MMP-cliffs, the following structural and potency criteria were applied. For a qualifying MMP, the difference in size of the exchanged fragments was limited to at most eight non-hydrogen atoms, and the maximal size of an exchanged fragment was set to 13 non-hydrogen atoms. Furthermore, the potency difference between compounds in an MMP meeting the structural criteria had to be at least 2 orders of magnitude. Thus, because we only considered compounds with at least 10 μM potency, at least one cliff forming compound was always potent in the nanomolar range, which we considered an additional requirement.

RESULTS AND DISCUSSION

MMP-Cliff Criteria. Similarity and potency criteria for the definition of MMP-cliffs were carefully considered. On the basis

dx.doi.org/10.1021/ci3001138 | J. Chem. Inf. Model. 2012, 52, 1138−1145

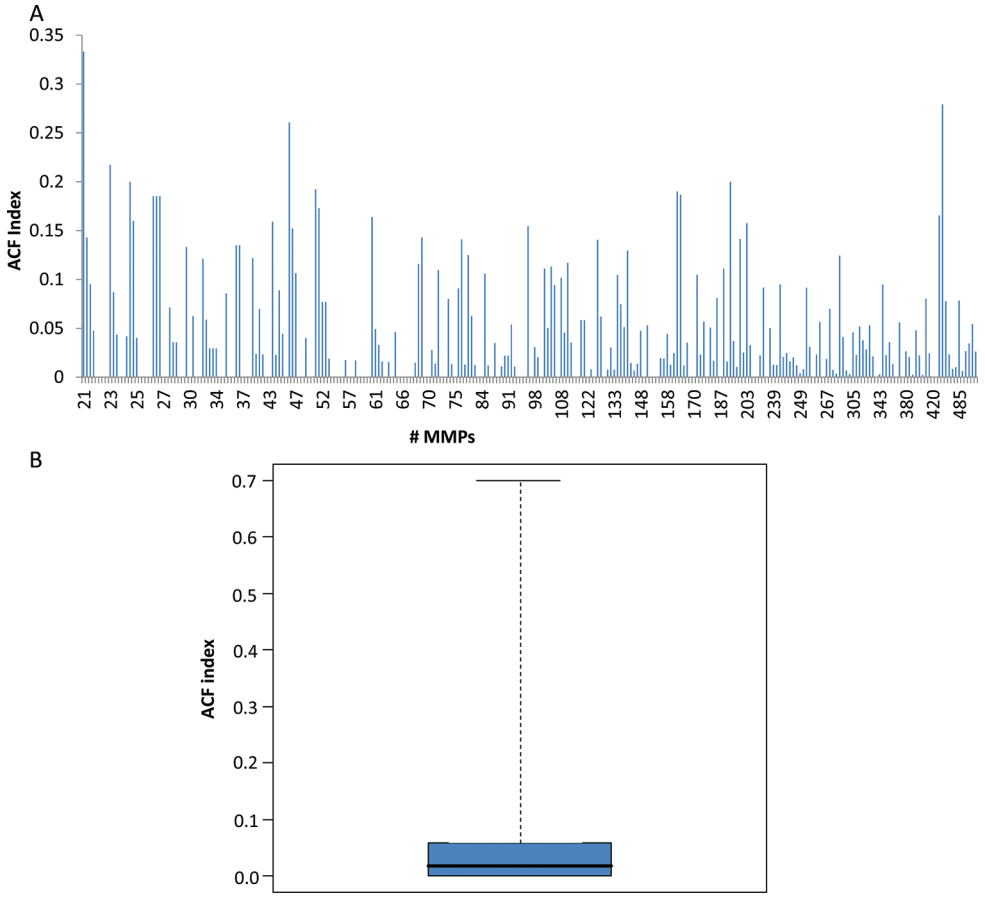


Figure 4. Activity cliff frequency index. (A) For target sets containing more than 20 and less than 500 MMPs, the activity cliff frequency (ACF) indices are reported. (B) The distribution of ACF indices for all target sets yielding MMPs is reported as a box plot. This representation provides the smallest ACF index (bottom line), lower quartile (lower boundary of the box), median value (thick line), upper quartile (upper boundary of the box), and the largest index (top line). The dashed line indicates the range.

of our potency criteria, the lower potency boundary of cliff formation was defined by two compounds with 10 μM and 100 nM potency, respectively. This was the "weakest" permitted cliff, and pairs of compounds with micromolar potency were not considered. As a criterion for structural similarity, we focused on substructure exchanges of only limited size, in accord with the activity cliff concept. Therefore, chemically intuitive upper transformation size boundaries were defined. Specifically, (i) transformations corresponding to the addition of a substituted six-membered ring (e.g., a phenol substituent) to a compound or (ii) the replacement of a five- or six-membered ring by a substituted condensated two-ring systems (containing a maximum of 10 ring atoms) were permitted as changes of maximal size. The latter transformation would allow, for example, the replacement of an imidazole with an indole ring or of a phenyl substituent with a condensated two-ring system. On the basis of these considerations, the size and size difference of exchanged fragments was limited to maximally 13 and eight non-hydrogen atoms, respectively. Largest permitted substructure exchanges for MMP-cliffs (and exemplary non-permitted substitutions) are illustrated in Figure 2. On the basis of these criteria, a systematic search for MMP-cliffs was performed for compounds active against more than 600 different targets.

MMP and Cliff Statistics. The 621 target sets yielded a total of 314,209 unique MMPs, and 21,285 of these pairs met our cliff criteria (Table 1). Thus, ~6.8% of all MMPs represented activity cliffs. The 21,285 MMP-cliffs involved a total of 13,980 active compounds (i.e., ~23.8%). Furthermore, 361 of our 621 target sets contained one or more MMP-cliffs (~58.1%) and 241 sets at least five cliffs (~38.8%). A complete listing of the 361 target sets containing at least 1 MMP-cliff is given in Table S1 of the Supporting Information. For all target sets, detailed compound data and activity cliff statistics are provided in Table S2 of the Supporting Information.

MMP-Cliff Distributions. We first determined the overall distribution of cliffs. The 361 target sets in which MMP-cliffs were detected contained on average ~63 cliffs. A total of 120 sets contained fewer than five cliffs. Figure 3A reports the top 30 target sets with the largest numbers of MMP-cliffs, all of which contained close to or more than 300 cliffs. Interestingly, 23 of the top 30 sets were directed against different G-protein coupled receptors. In six cases, more than 600 cliffs were found, with factor Xa and thrombin inhibitors yielding very large numbers of more than 1800 and 1300 MMP-cliffs, respectively. Figure 3B shows the number of MMP-cliffs as a function of the number of compounds per target set. As expected, the number of MMP-cliffs often, but not always, increased with the number of compounds. The number of MMP-cliffs was then normalized with respect to the number of MMPs formed per target set by calculating an activity cliff frequency (ACF) index, defined as the number of cliffs divided by the total number of MMPs per set. Figure 4A shows a representative sample of ACF indices for target sets that yielded between 20 and 500 MMPs. In addition,

dx.doi.org/10.1021/ci3001138 | J. Chem. Inf. Model. 2012, 52, 1138−1145

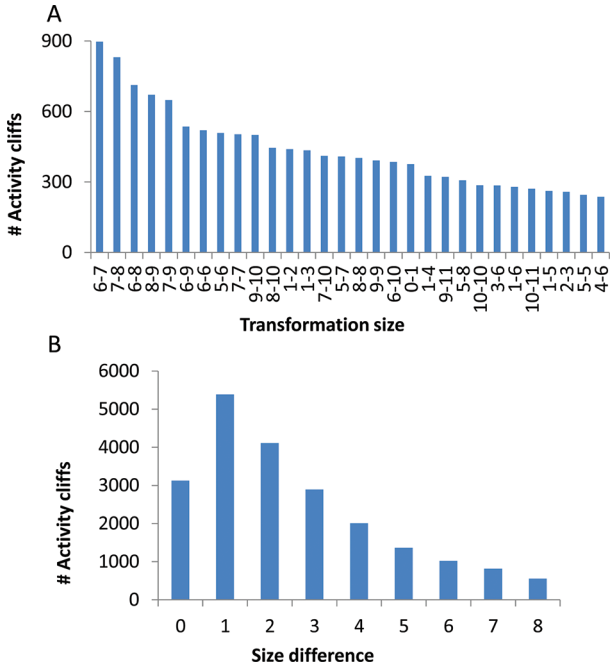


Figure 5. Distribution of transformation sizes. (A) The number of MMP-cliffs is reported for the top 30 most frequent fragment size combinations. For instance, ~900 cliffs involve the exchange of substructures consisting of six and seven (non-hydrogen) atoms, respectively (combination 6−7). (B) Shown is the distribution of MMP-cliffs involving transformations of equally or differently sized substructures, with size differences ranging from 0 to 8.

Table 2. Tanimoto Similarity of MMP-Cliffs

MACCS Tc	# MCs	Cliff Ratio
0.75	17,152	81%
0.80	14,433	68%
0.85	10,249	48%

The numbers of MMP-cliffs (# MCs) and their percentage (Cliff Ratio) exceeding MACCS Tanimoto coefficient (Tc) values of 0.75, 0.80, or 0.85 are reported. For example, the Cliff Ratio of 48% means that compounds forming only 48% of all MMP-cliffs reach or exceed the level of 85% MACCS Tanimoto similarity.

Figure 4B shows a box plot representation of ACF indices for all target sets containing MMPs. Although the distribution of MMPs greatly varied over different target sets, MMP-cliff propensities were generally low, and ACF indices were similar for more than 75% of the target sets, indicating a similar frequency of occurrence of MMP-cliffs for many target sets.

Cliff Transformations. We next analyzed the sizes of substructures exchanged between cliff forming compounds. Figure 5A reports the top 30 most frequent fragment size combinations. Substructures consisting of five to eight atoms were most frequently substituted. However, small transformations involving only one to three atoms (including the addition of one or two non-hydrogen atoms) were also often observed. At the other end of the permitted fragment size range, substructures containing 10 atoms were frequently exchanged. In Figure 5B, the distribution of substructure size differences is reported. It reveals that the majority of fragments exchanged between cliff forming compounds were of very similar size. Differences of only one or two atoms were most frequently observed, indicating the presence of only limited structural changes in many MMP-cliffs.

Representative MMP-Cliffs. In Figure 6, representative MMP-cliffs are shown for all permitted differences in the size of exchanged substructures (0−8). By design, substructure exchanges in MMP-cliffs were well-defined and restricted to single sites. Modifications often involved R-groups, but also changes in a core structure region. The locally confined nature of structural modifications revealed by MMP-cliffs made it straightforward to assess these chemical changes. As illustrated in Figure 6, frequently exchanged substructures included differently substituted rings, small fragments, and individual atoms, as already indicated by the preferred transformation sizes we observed. Taken together, these observations supported the choice of our criteria for permitted ranges of fragment sizes and size differences. Of course, depending on the application, permitted fragments size and size difference ranges can be easily modified.

Similarity Calculations for MMP-Cliffs. As stated above, we found that ~6.8% of all accepted MMPs and ~23.8% of all active compounds that participated in MMPs yielded MMP-cliffs. In a previous analysis of ChEMBL/BindingDB, activity cliffs were defined on the basis of Tanimoto similarity calculations using MACCS structural keys[19] as descriptors, and the presence of 85% MACCS Tanimoto similarity was required as a cliff criterion.[17] Compared to our current analysis, similar but distinct measurement and activity cliff criteria were applied. In particular, compounds falling within the potency range from 100 nM to 10 μM were excluded from activity cliff formation. It was found that ~2% of all qualifying compound pairs formed activity cliffs spanning at least 2 orders of magnitude in potency, and that these cliffs involved ~10% of all active compounds that were analyzed.[17] Because the former and current analyses are not directly comparable, we also calculated MACCS Tanimoto similarity for the MMP-cliffs we identified. The results are reported in Table 2. For only 48% of our MMP-cliffs, a MACCS Tc value of at least 0.85 was obtained. Thus, only about half of these MMP-cliffs would have been detected on the basis of calculated similarity values. Figures 7 and 8 show examples of activity cliffs that were conserved on the basis of Tanimoto similarity and MMP criteria or only identified as MMP-cliffs, respectively. Both conserved and unique MMP-cliffs were generally members of analog series and as such qualified as cliffs from a chemical point of view. Hence, unique MMP-cliffs shown in Figure 8 also illustrate limitations of whole-molecular similarity calculations for activity cliff assignments, especially for relatively small compounds and chemically different substituents. In many instances, pairs of structural analogs were too dissimilar on the basis of Tanimoto calculations to be considered activity cliffs.

■ CONCLUSIONS

In this study, we have introduced MMP-cliffs as a new and entirely substructure-based representation of activity cliffs. In addition, for the formation of MMP-cliffs, only equilibrium potency measurements were considered. In a large-scale compound data mining effort, MMP-cliffs were systematically identified for compounds active against more than 600 targets. The locally confined nature of chemical transformations defining MMP-cliffs and their restriction to a single site are attractive features of the approach and naturally guide the search for activity cliffs toward analog series. Transformation size restrictions can be easily introduced and modified to focus the analysis on specific subsets of MMP-cliffs. We found that

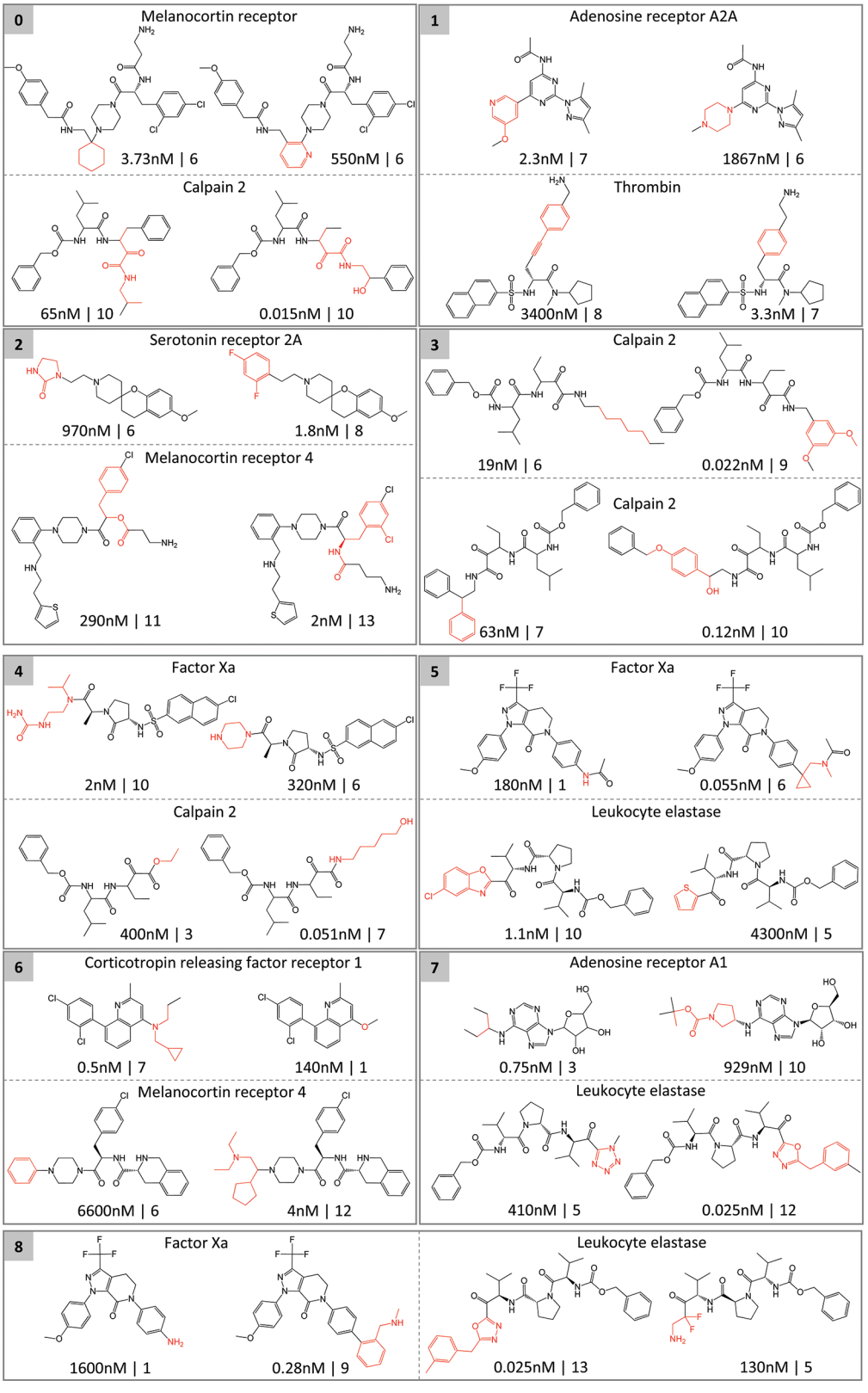


Figure 6. Representative MMP-cliffs. Exemplary cliffs are shown with size differences between exchanged fragments ranging from zero to eight non-hydrogen atoms. For each size difference (0−8), two representative activity cliffs are shown. The exchanged substructures are colored red. For each cliff, the target name, compound potency values, and the size of the exchanged substructures are reported. For example, "3.73nM | 6" indicates that the potency value of the compound is 3.73 nM and the number of non-hydrogen atoms comprising the differentiating fragment is 6.

most of the transformations leading to cliffs involved only small substructures of similar size (containing five to eight atoms). The introduction of MMP-cliffs provides an alternative to the identification of activity cliffs using conventional molecular similarity calculations. Substructure-based approaches do not make such similarity calculations obsolete, but help to evaluate compound

dx.doi.org/10.1021/ci3001138 | J. Chem. Inf. Model. 2012, 52, 1138−1145

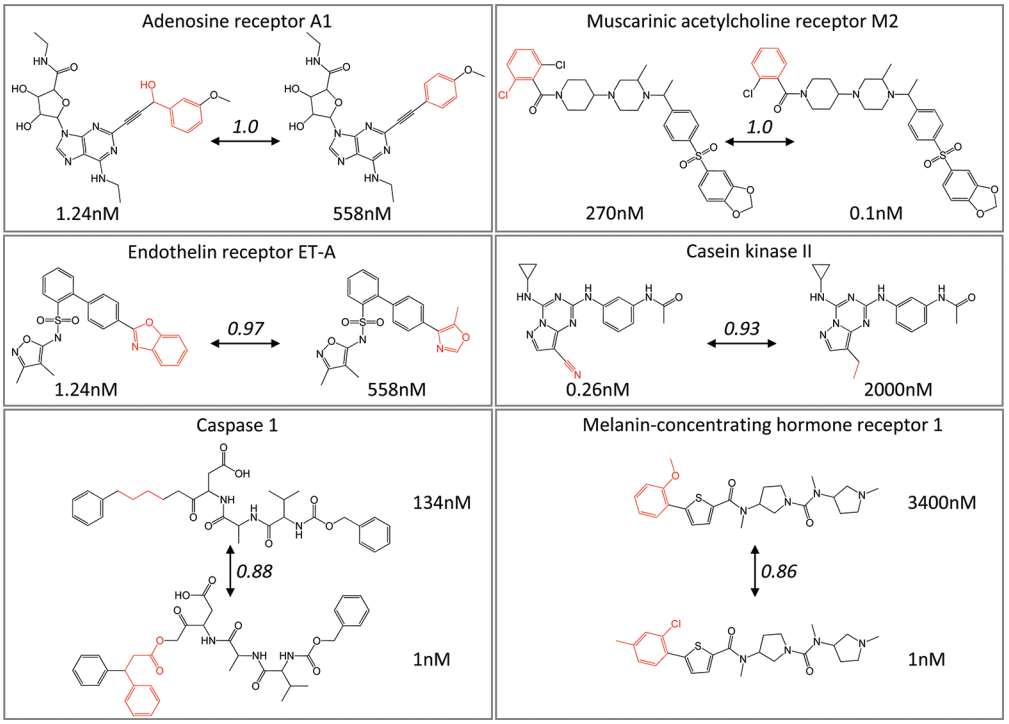


Figure 7. Conserved activity cliffs. Shown are six exemplary MMP-cliffs for which the cliff partners exceed a MACCS Tanimoto coefficient (Tc) value of 0.85 (given in italics). In each case, the exchanged fragments are colored red and the name of the target and compound potency values are given.

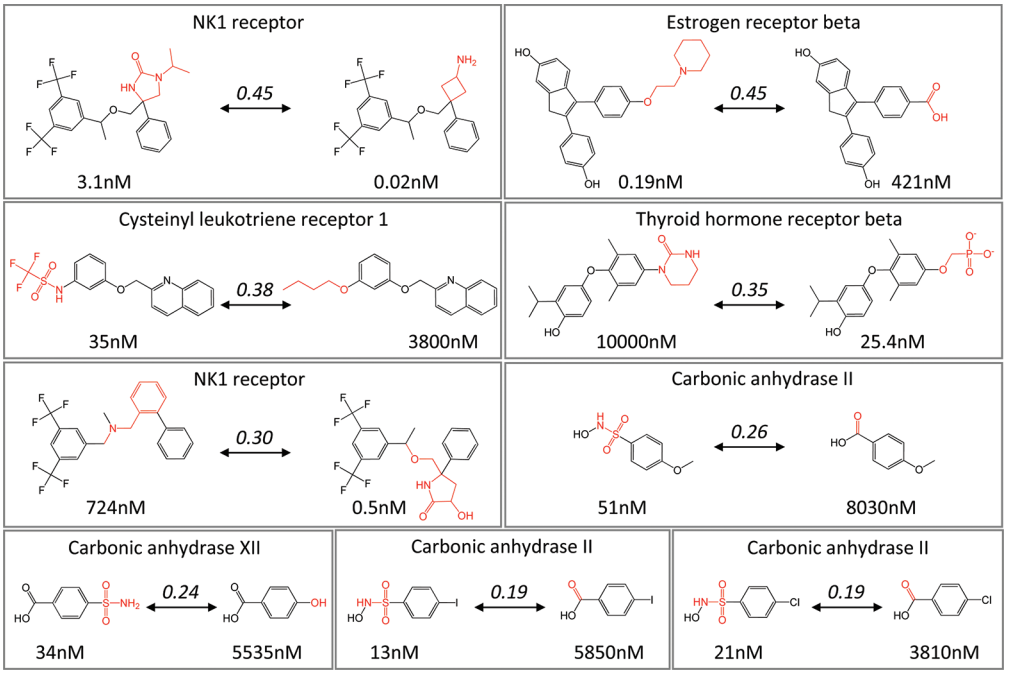


Figure 8. Unique MMP-cliffs. Nine exemplary MMP-cliffs are shown for which low MACCS Tc values of ≤0.45 were obtained. The representation is according to Figure 7.

similarity from different points of view. Approximately 20−50% of the MMP-cliffs that were identified could not be reproduced on the basis of Tanimoto similarity calculations, depending on the similarity threshold values that were applied. Hence, MMP-cliffs provide an alternative source of activity cliff information and are likely to yield more candidates for further chemical exploration than the application of conventional similarity measures. Because similarity calculations are capable of detecting analogous compounds that are modified at more than one site, the activity cliff information provided by substructure and whole-molecule similarity approaches will often be complementary.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Table S1 and S2 report the target set, compound data, and activity cliff statistics. This information is available free of charge via the Internet at http://pubs.acs.org.

dx.doi.org/10.1021/ci3001138 | J. Chem. Inf. Model. 2012, 52, 1138−1145

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

X.H. is supported by the China Scholarship Council and D.S. 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*, DOI: 10.1021/jm201706b.
(3) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity landscape representations for structure−activity relationship analysis. *J. Med. Chem.* **2010**, *53*, 8209−8223.
(4) Guha, R.; Van Drie, J. H. Structure−activity landscape index: Identifying and quantifying activity cliffs. *J. Chem. Inf. Model* **2008**, *48*, 646−658.
(5) 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.
(6) 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.
(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) Willett, P. Searching techniques for databases of two- and three-dimensional structures. *J. Med. Chem.* **2005**, *48*, 4183−4199.
(9) Wawer, M.; Bajorath, J. Local structural changes, global data views: Graphical substructure−activity relationship trailing. *J. Med. Chem.* **2011**, *54*, 2944−2951.
(10) 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.
(11) 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.
(12) 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.
(13) ChEMBL. http://www.ebi.ac.uk/chembldb/ (accessed February 8, 2012).
(14) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A Web-accessible database of experimentally determined protein−ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198−D201.
(15) BindingDB. http://www.bindingdb.org/ (accessed February 8, 2012).
(16) PubChem BioAssay. http://pubchem.ncbi.nlm.nih.gov/ (accessed February 8, 2012).
(17) 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.
(18) 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.
(19) *MACCS Structural Keys*; Symyx Software: San Ramon, CA, 2005.