JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

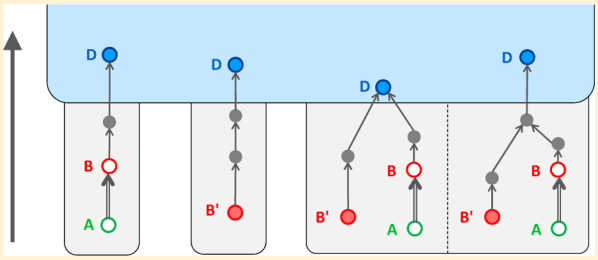
# Compound Pathway Model To Capture SAR Progression: Comparison of Activity Cliff-Dependent and -Independent Pathways

Dagmar Stumpfe, Dilyana Dimova,[†] Kathrin Heikamp,[†] 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:** A compound pathway model is introduced to monitor SAR progression in compound data sets. Pathways are formed by sequences of structurally analogous compounds with stepwise increasing potency that ultimately yield highly potent compounds. Hence, the model was designed to mimic compound optimization efforts. Different pathway categories were defined. Pathways originating from any active compound in a data set were systematically identified including compounds forming activity cliffs. The relative frequency of activity cliff-dependent and -independent pathways was determined and compared. In 23 of 39 different compound data sets that qualified for our analysis, significant differences in the relative frequency of activity cliff-dependent and -independent pathways were observed. In 17 of these 23 data sets, activity cliff-dependent pathways occurred with higher relative frequency than cliff-independent pathways. In addition, pathways originating from the majority of activity cliff compounds displayed desired SAR progression, reflecting SAR information gain associated with activity cliffs.




## INTRODUCTION

Capturing SAR information in compound data sets of any source is a prime task in computational medicinal chemistry,[1,2] in addition to, for example, QSAR-based compound activity predictions.[3] SAR information can be extracted from compound activity data through numerical and/or graphical analysis.[1,2] In general terms, SAR information extraction requires the systematic study of structural and potency relationships between active compounds and the identification of series with defined SAR characteristics such as SAR continuity or discontinuity.[2] Numerical and graphical analysis components are typically combined to model activity landscape representations of compound data sets that integrate structural and potency relationships in a systematic manner.[2] Cardinal features of activity landscapes are activity cliffs that are formed by pairs or groups of compounds with large potency differences.[4−6] The activity cliff concept is popular because activity cliffs encode small structural changes leading to large potency alterations, which is generally thought to associate activity cliff compounds with high SAR information content.[4,6] Activity landscapes and activity cliffs have been studied using a variety of molecular representations and similarity measures[2,6−8] also including structurally conservative approaches that limit the formation of cliffs to analogous compounds.[9] Compared to whole-molecule similarity calculations, the latter approach further supports the interpretation of activity cliffs from a chemical perspective, which is important for practical SAR analysis. Systematic surveys of activity cliffs in public domain compounds have been carried out to characterize activity cliff populations[10,11] and their potency range distributions.[11] Depending on the chosen molecular representations and similarity criteria considered, only ∼4−6% of all pairs of qualifying similar active compounds formed activity cliffs, confirming that cliffs are a rare activity landscape feature. On the other hand, ∼20−30% of all compounds with high-confidence activity annotations across different targets were found to be involved in the formation of at least one large-magnitude activity cliff.[11] Thus, activity cliffs can be identified in essentially all compound data sets and provide possible starting points for SAR exploration.

Although activity cliffs have been studied in different ways, either from a more computational or chemical perspective, they have thus far not been evaluated in the context of SAR progression, which refers to compound series with steadily increasing potency that ultimately yield highly potent compounds. Therefore, we put the evaluation of activity cliffs here into the broader context of compound pathway analysis. Introducing a pathway model that mimics chemical optimization, we systematically monitored SAR progression in a variety of data sets evolving over time by considering any compound as a potential starting point including compounds forming activity cliffs. Overall, activity cliff-dependent pathways leading to highly potent compounds were observed with higher relative frequency than pathways originating from other active compounds. These findings further support the preferential consideration of activity cliffs for SAR exploration and compound optimization.

**Received:** March 5, 2013

 © XXXX American Chemical Society  dx.doi.org/10.1021/ci400141w | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

## MATERIALS AND METHODS

**Transformation Size-Restricted Matched Molecular Pairs.** A matched molecular pair (MMP)[12] is formed by two structurally related compounds that are distinguished at a single site through the exchange of a substructure, a so-called chemical transformation.[13] Transformation size restrictions have been introduced to confine MMPs to structurally analogous compounds distinguished by functional groups or an individual ring system.[9] Such transformation size-restricted MMPs were calculated using an in-house implementation of the algorithm by Hussain and Rea.[13] If several transformations met the size limitations for a given compound pair, the smallest transformation was selected.

**Activity Cliff Criteria.** In our analysis, we followed the definition of MMP cliffs,[9] a structurally conservative approach that usually limits the formation of cliffs to structural analogs, which we considered important in the context of compound pathway analysis. Accordingly, as a similarity criterion, the formation of a transformation size-restricted MMP was applied, and as a potency difference criterion, a difference in equilibrium constants between the two MMP-forming compounds of at least two orders of magnitude was required.

**Potency-Directed Compound Pathways.** MMP-based potency-directed compound pathways were introduced to capture compound series with positive potency progression. Pathway compounds were required to form stepwise overlapping MMPs, e.g., three compounds X, Y, and Z qualified for a pathway X−Y−Z if the two MMPs [X,Y] and [Y,Z] existed. In addition, pathway compounds were required to have stepwise increasing potency (i.e., potency X < Y < Z). Furthermore, the endpoint of a pathway (i.e., Z) had to belong to the 10% most potent compounds within a data set (in the following referred to as *D*), and the starting point (i.e., X) had to fall outside of this subset of the most potent compounds.

Two different categories of pathways were distinguished depending on their origin. In a given compound data set, all activity cliffs were determined, and each highly potent activity cliff partner not belonging to the top 10% most potent compounds, in the following referred to as *B*, provided a potential starting point for *activity cliff-dependent compound pathways*. Weakly potent activity cliff partners (referred to as *A*) were not considered as pathway starting points. The activity cliff compound defined the beginning of the time course for pathway progression. In addition, all compounds in a data set not involved in the formation or progression of any activity cliff and not belonging to the top 10% most potent compounds, in the following referred to as *C*, provided potential starting points for *activity cliff-independent compound pathways*. All pathways formed by a minimum of two compounds including the starting point were calculated. The pathway model is schematically illustrated in Figure 1. Increasingly potent pathway candidate compounds were only considered if they became available during the same or subsequent years compared to the preceding pathway compound (i.e., a more potent analog was not included in a pathway if it became available earlier than the preceding compound). Thus, pathway analysis monitored possible SAR progression over a time course.

Compound data analysis and pathway calculations were carried out with in-house generated Java programs or KNIME[14] protocols.

**Pathway Frequency.** A normalized pathway frequency was calculated for activity cliff-dependent and -independent compound pathways. Therefore, for each of the top 10% most potent compounds (*D* in Figure 1A), the ratio of qualifying pathways originating from *B* (i.e., activity cliff-dependent) or *C* compounds (i.e., activity cliff-independent) over all possible compound pathways was calculated (i.e., all possible MMP sequences leading from *B* or *C* to *D*, respectively). For each data set, the average normalized frequencies were calculated for all *D* compounds that were pathway endpoints.

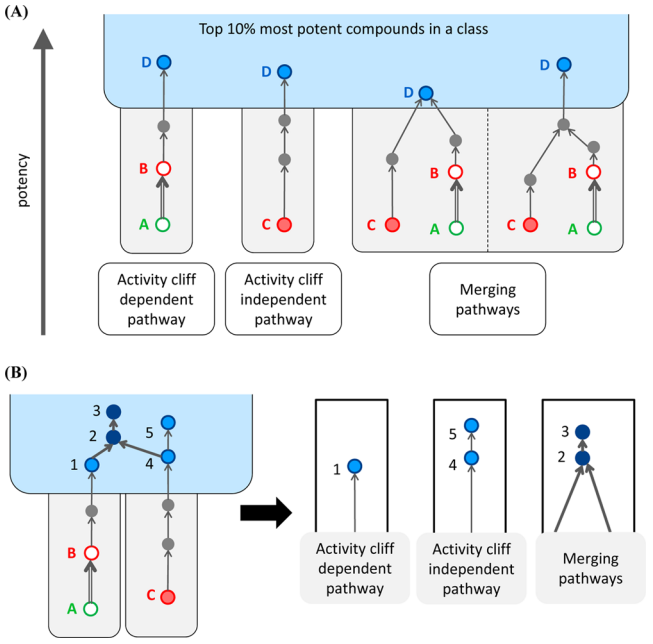



**Figure 1.** Compound pathways. (A) Schematic illustration of different categories of qualifying potency-directed compound pathways. Highly potent activity cliff partners (*B* compounds) represent potential starting points of activity cliff-dependent pathways. The weakly potent cliff partners are denoted as *A* compounds and not further considered. Data set compounds not involved in a cliff formation and not a part of a cliff-dependent pathway (*C* compounds) represent potential starting points of activity cliff-independent pathways. Activity cliff-dependent and -independent pathways can merge either within the subsets of the 10% most potent data set compounds (*D*) or prior to reaching these compounds. Intermittent pathway compounds are depicted in gray. (B) *D* compounds belonging to different types of pathways. Compound 1 belongs only to an activity cliff-dependent pathway. Compounds 2 and 3 belong to a merging pathway. Compounds 4 and 5 belong only to an activity cliff-independent pathway.

**Compound Data Sets.** To evaluate SAR progression in a systematic manner, we searched ChEMBL[15] (release 14) for target-based compound data sets that evolved over time and in which at least a 10% average frequency for activity cliff-dependent or -independent pathways was observed. Each data set had to contain at least 100 compounds active against a human target with direct interactions (ChEMBL relationship type "D") at the highest confidence level (ChEMBL confidence score "9").[15] In addition, equilibrium constants ($K_i$ values) had to be available as potency measurements for all data set compounds. If several $K_i$ values were available for a compound, the most recent measurement was used. To account for data set evolution over time, the condition was applied that compounds comprising a set had to be reported in increments over a period of at least five subsequent years. During each year, the addition of a new compound subset was required.


dx.doi.org/10.1021/ci400141w | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

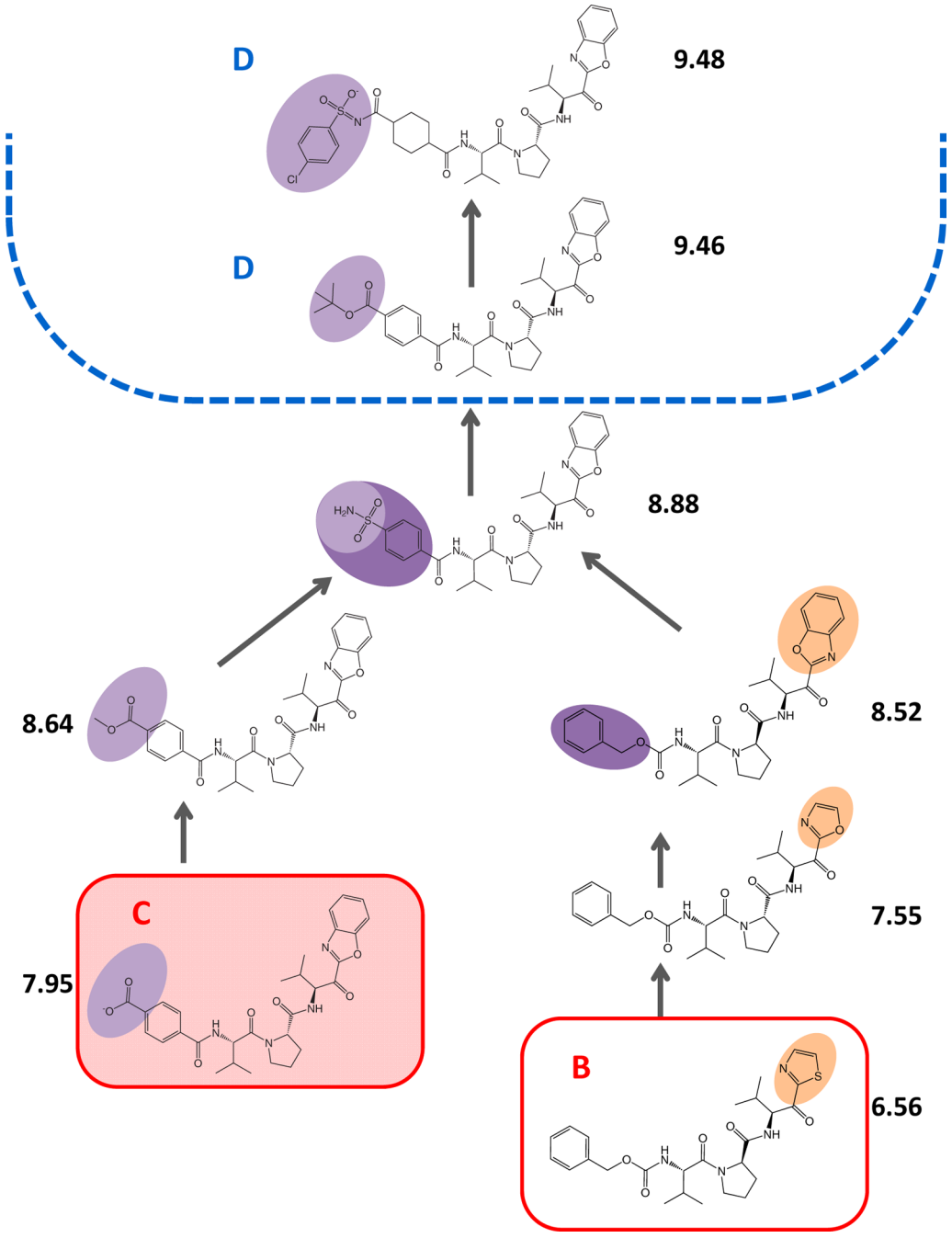



**Figure 2.** Exemplary pathway. Example of a merging pathway from the leukocyte elastase inhibitor data set shown in detail. Pathway starting points are color-coded according to Figure 1. Structural changes between pairs of compounds in a pathway are highlighted using corresponding colors, and compounds are labeled with their p$K_i$ values.

## RESULTS AND DISCUSSION

**Pathway Model of SAR Progression.** Our compound pathway model was designed to capture SAR progression in a compound activity class. Series of structurally analogous compounds with steady potency progression toward the most potent compounds found in a data set served as a model for compound optimization paths. Therefore, structural relationships between compounds were established on the basis of sequentially overlapping MMPs with size-restricted structural changes. Furthermore, the formation of pathways was limited to series of increasingly potent compounds that were reported during the same or subsequent years. Hence, a more potent analog in a pathway was not permitted to be reported earlier than a less potent one. This additional restriction was introduced to model compound optimization over time. Pathways were differentiated according to their origins and potential overlap, as further discussed below.

**Activity Cliff-Dependent and -Independent Pathways.** Applying the pathway model, we systematically determined qualifying pathways that originated from active compounds and compared activity cliff-dependent and -independent pathways. Figure 1A illustrates the different pathway categories we considered. Activity cliff-dependent pathways start from the highly potent activity cliff partner and contain compounds with further increasing potency. The assumption underlying the


dx.doi.org/10.1021/ci400141w | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

**Table 1. Evolving Compound Data Sets and Pathway Statistics**[a]

| no. | ChEMBL target ID | target | #Cpds | #*D* from only *B* | #*D* from only C | #*D* from *B* and *C* | *B* norm. freq. | *C* norm. freq. | Δ norm. freq. |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 4617 | phenylethanolamine N-methyltransferase | 148 | 10 | 0 | 1 | 0.85 | 0.08 | 0.77 |
| 2 | 249 | neurokinin 1 receptor | 211 | 7 | 1 | 0 | 0.88 | 0.17 | 0.71 |
| 3 | 299 | protein kinase C alpha | 168 | 9 | 2 | 2 | 0.76 | 0.17 | 0.59 |
| 4 | 2243 | anandamide amidohydrolase | 101 | 1 | 6 | 0 | 1.00 | 0.43 | 0.58 |
| 5 | 248 | leukocyte elastase | 192 | 0 | 7 | 7 | 0.93 | 0.48 | 0.45 |
| 6 | 3795 | melanocortin receptor 1 | 134 | 5 | 0 | 0 | 0.38 | 0.00 | 0.38 |
| 7 | 211 | muscarinic acetylcholine receptor M2 | 199 | 5 | 0 | 0 | 0.36 | 0.00 | 0.36 |
| 8 | 3837 | cathepsin L | 201 | 11 | 2 | 0 | 0.40 | 0.13 | 0.27 |
| 9 | 268 | cathepsin K | 272 | 18 | 0 | 0 | 0.26 | 0.00 | 0.26 |
| 10 | 1997 | equilibrative nucleoside transporter 1 | 118 | 4 | 0 | 0 | 0.25 | 0.00 | 0.25 |
| 11 | 1855 | gonadotropin-releasing hormone receptor | 267 | 12 | 5 | 5 | 0.33 | 0.08 | 0.24 |
| 12 | 4822 | beta-secretase 1 | 116 | 2 | 0 | 0 | 0.22 | 0.00 | 0.22 |
| 13 | 5071 | G protein-coupled receptor 44 | 376 | 18 | 8 | 5 | 0.35 | 0.15 | 0.20 |
| 14 | 3759 | histamine H4 receptor | 334 | 13 | 6 | 7 | 0.32 | 0.13 | 0.19 |
| 15 | 4561 | neuropeptide Y receptor type 5 | 234 | 11 | 8 | 2 | 0.31 | 0.16 | 0.15 |
| 16 | 213 | beta-1 adrenergic receptor | 175 | 7 | 3 | 1 | 0.30 | 0.18 | 0.12 |
| 17 | 344 | melanin-concentrating hormone receptor 1 | 870 | 32 | 18 | 28 | 0.20 | 0.09 | 0.11 |
| 18 | 1889 | vasopressin V1a receptor | 318 | 13 | 0 | 7 | 0.41 | 0.32 | 0.09 |
| 19 | 2954 | cathepsin S | 371 | 14 | 6 | 2 | 0.19 | 0.11 | 0.08 |
| 20 | 2014 | nociceptin receptor | 599 | 14 | 17 | 12 | 0.23 | 0.17 | 0.06 |
| 21 | 4308 | bradykinin B1 receptor | 415 | 19 | 6 | 3 | 0.13 | 0.08 | 0.05 |
| 22 | 219 | dopamine D4 receptor | 436 | 8 | 21 | 0 | 0.18 | 0.15 | 0.03 |
| 23 | 244 | coagulation factor X | 1198 | 25 | 32 | 26 | 0.14 | 0.11 | 0.03 |
| 24 | 245 | muscarinic acetylcholine receptor M3 | 343 | 10 | 13 | 0 | 0.28 | 0.25 | 0.02 |
| 25 | 234 | dopamine D3 receptor | 881 | 26 | 25 | 12 | 0.12 | 0.10 | 0.02 |
| 26 | 214 | serotonin 1a (5-HT1a) receptor | 938 | 16 | 49 | 2 | 0.14 | 0.12 | 0.02 |
| 27 | 1800 | corticotropin releasing factor receptor 1 | 477 | 14 | 15 | 9 | 0.26 | 0.24 | 0.01 |
| 28 | 228 | serotonin transporter | 1099 | 21 | 63 | 3 | 0.14 | 0.13 | 0.01 |
| 29 | 204 | thrombin | 808 | 23 | 21 | 13 | 0.12 | 0.12 | 0.00 |
| 30 | 259 | melanocortin receptor 4 | 1273 | 51 | 21 | 45 | 0.09 | 0.10 | −0.01 |
| 31 | 231 | histamine H1 receptor | 187 | 1 | 9 | 0 | 0.33 | 0.38 | −0.04 |
| 32 | 3371 | serotonin 6 (5-HT6) receptor | 888 | 3 | 54 | 1 | 0.02 | 0.12 | −0.10 |
| 33 | 3798 | calcitonin gene-related peptide type 1 receptor | 246 | 2 | 16 | 0 | 0.11 | 0.21 | −0.10 |
| 34 | 284 | dipeptidyl peptidase IV | 276 | 8 | 13 | 0 | 0.21 | 0.38 | −0.17 |
| 35 | 1836 | prostanoid EP4 receptor | 194 | 3 | 9 | 0 | 0.08 | 0.28 | −0.20 |
| 36 | 264 | histamine H3 receptor | 1515 | 9 | 94 | 8 | 0.06 | 0.28 | −0.22 |
| 37 | 1945 | melatonin receptor 1A | 215 | 0 | 11 | 0 | 0.00 | 0.29 | −0.29 |
| 38 | 1946 | melatonin receptor 1B | 262 | 1 | 16 | 0 | 0.04 | 0.42 | −0.38 |
| 39 | 1914 | butyrylcholinesterase | 159 | 0 | 13 | 0 | 0.00 | 0.87 | −0.87 |

[a]All 39 compound data sets meeting the selection criteria are listed. For each set, the number of compounds (#Cpds) and the ChEMBL target ID are given. In addition, the numbers of *D* compounds detected only by activity cliff-dependent pathways (#*D* from only *B*), only by activity cliff-independent pathways (#*D* from only *C*), or by merging pathways (#*D* from *B* and *C*) are reported. Furthermore, the normalized frequencies of activity cliff-dependent (*B* norm. freq.) and -independent (*C* norm. freq.) and the frequency difference (Δ norm. freq.) are given. Data sets are ranked in the order of decreasing frequency difference.

activity cliff concept is that comparison of the cliff partners provides interpretable SAR information and identifies potential SAR determinants that might aid in the design of compounds with further increased potency. For activity cliff-independent pathways that can originate from any other data set compound, no such SAR information is available at the beginning. In addition, merging pathways can be found that combine compounds from activity cliff-dependent and -independent pathways either before or after reaching the top 10% most potent compounds. Figure 1B shows *D* compounds dependent on their pathway membership(s). For statistical analysis, merging pathways are counted as both activity cliff-dependent and -independent pathways because their *D* compounds can be separately reached by both pathway categories. Figure 2 shows an exemplary merging pathway in detail.

**Pathway Detection.** Key questions of our analysis included whether (i) the pathway model would reveal differences in the distribution of different compound pathway categories and whether or not (ii) SAR information associated with activity cliffs might more frequently result in compound pathways with SAR progression than the use of other active compounds as starting points.


dx.doi.org/10.1021/ci400141w | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Therefore, we have systematically determined all activity cliff-dependent, -independent, and merging pathways in the 39 evolving compound data sets.

Table 1 reports that differences in relative pathway frequency of different magnitude were indeed observed in many data sets. In a few data sets, activity cliff-dependent or -independent pathways originated on average from nearly each *B* or *C* compounds, respectively, whereas in other sets one and/or the other pathway category occurred with only very low frequency, as further discussed below. The presence of compound pathways represented a general diagnostic of SAR information content in the evolving compound data sets. Substantial differences in SAR information content were observed on the basis of pathway frequency.

Concerning pathway statistics, the following general criteria were taken into consideration. On average, there were 8.9 and 45.6 *B* and *C* compounds per data set, respectively, which met the starting point criteria. Because the data sets contained more qualifying *C* than *B* compounds, there was an intrinsically higher statistical probability to observe cliff-independent pathways. On the other hand, B compounds as pathway starting points would have a higher likelihood to reach the most potent *D* compounds in a data set provided the potency of *B* compounds was generally higher than the potency of (noncliff) *C* compounds. This possibility was examined by comparing the potency distribution of *B* and *C* compounds across all data sets, as reported in Figure 3. The results show that the potency

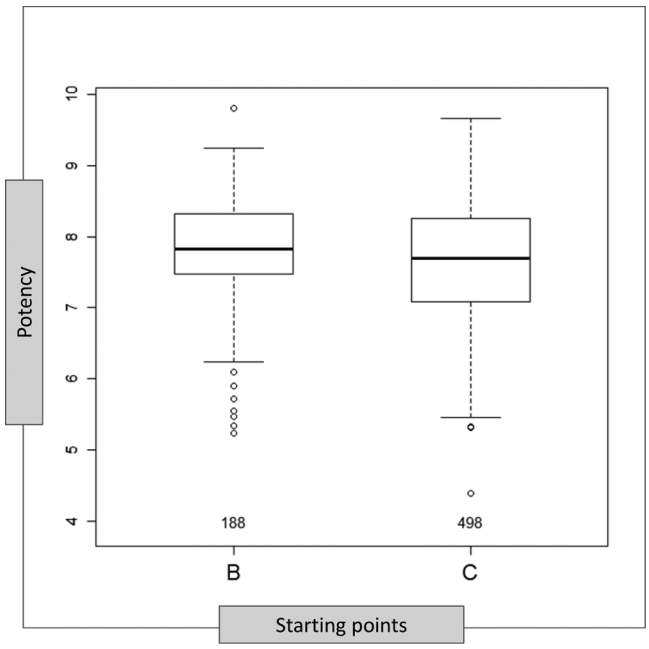



**Figure 3.** Potency distribution of pathway starting points. Potency distribution of the pathway starting points *B* and *C* from all data sets in boxplots.

distribution and median potency values for starting points of activity cliff-dependent and -independent pathways were very similar. Hence, there was no significant potency level advantage for pathways originating from activity cliffs.

**Pathway Comparison.** Given the general differences in the number of *B* and *C* compounds, pathway utilization was quantified and compared by calculating the relative frequency of activity cliff-dependent and -independent pathways with SAR progression normalized with respect to the number of all possible pathways within each category, i.e., all MMP sequences originating from either *B* or *C* compounds. As rationalized above, for statistical analysis, merging pathways qualified as both cliff-dependent and -independent pathways. The results of pathway frequency calculation and comparison reported in Table 1 reveal differences in the normalized frequency of pathways originating from *B* and *C* compounds of more than 10% in 23 of 39 data sets. In 16 data sets, the pathway frequencies were comparable, although their magnitude varied considerably (reflecting data set-dependent differences in SAR information content). The frequency difference distribution for the 23 data sets with more than 10% difference is monitored in Figure 4. In 17 of these 23 data sets, activity cliff-dependent

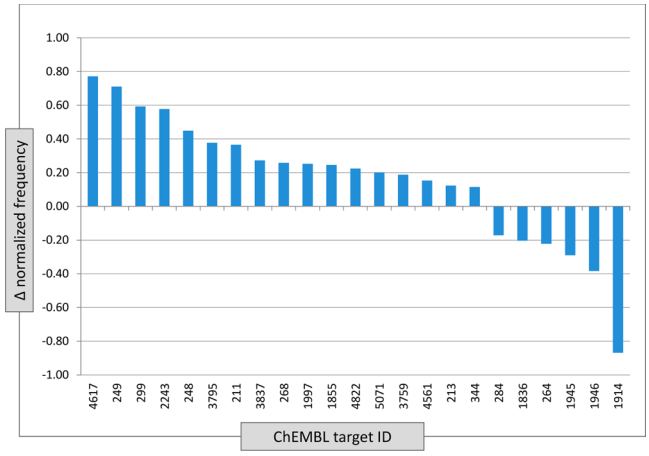



**Figure 4.** Pathway frequency difference. Difference between normalized frequencies of activity cliff-dependent and -independent pathways targets reported for all 23 data sets with a more than 10% difference. Positive and negative differences indicate larger frequencies of activity cliff-dependent and -independent pathways, respectively.

pathways were observed with in part much higher relative frequency than cliff-independent pathways. For example, for phenylethanolamine N-methyltransferase inhibitors, there were on average 0.85 qualifying pathways per *B* and close to 0 pathways per *C* compound. On the other end of the pathway frequency spectrum, the butyrylcholinesterase inhibitor data set stood out in which the relative frequency of qualifying pathways originating from C compounds was 0.87, while no activity cliff-dependent pathways were observed (despite the presence of activity cliffs in this data set). However, overall there were only six of 23 data sets in which activity cliff-independent pathways occurred with higher frequency, as shown in Figure 4.

In addition to pathway frequencies, we also determined which *D* compounds were reached by pathways belonging to different categories, as also reported in Table 1. On average, 72.0% of all *D* compounds in a data set were reached by pathways. In 21 of 39 cases, activity cliff-dependent pathways detected more *D* compounds than cliff-independent pathways, although on average five to six times more *C* than *B* compounds were available per data set. Furthermore, pathways originating from 53.9% and 28.1% of all *B* and *C* compounds, respectively, reached *D* compounds. Table S1 of the Supporting Information reports statistics of *B*, *C*, and *D* compounds and their pathway engagement for all data sets.

**Concluding Remarks.** We have investigated a compound pathway model to systematically detect compound series with


dx.doi.org/10.1021/ci400141w | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

SAR progression in diverse evolving data sets. Hence, pathway and SAR progression were monitored over time. Different pathway categories were defined to distinguish between compound pathways originating from activity cliffs and pathways originating from other active compounds. Activity cliffs present in all data sets were identified, and all qualifying activity cliff-dependent, -independent, and merging pathways were determined. The pathway model does not reveal how compounds are chemically explored and how activity cliffs are generated that serve as pathway start points. In fact, optimization efforts might produce compounds forming multiple and overlapping cliffs as a part of a variety of compound series. In our analysis, all activity cliffs were individually considered as potential pathway origins at the level of compound pairs. Compound pathways were required to follow a time course such that they also might represent optimization paths. Yet, it was not possible to determine on the basis of our analysis whether or not pathways represented actual optimization paths. However, the pathway model is designed to determine in a consistent manner where compound series with defined potency progression originate that lead to the most potent compounds present in a data set, and activity cliff-dependent and -independent pathways are clearly distinguished.

A key finding of our analysis has been that activity cliff-dependent pathways with desirable SAR progression were detected with higher relative frequency among potential paths than cliff-independent ones. Furthermore, pathways originating from the majority of activity cliffs reached highly potent compounds. Hence, there has been evidence for better SAR progression originating from activity cliffs than other compounds, consistent with the assumption that activity cliffs often reveal SAR determinants.

Hence, taken together, our findings supported the utility of the pathway model to monitor SAR progression in compound data sets and indicated that activity cliff-dependent pathways were more likely to yield well-defined SAR progression than pathways originating from other active compounds.

## ■ ASSOCIATED CONTENT

### Supporting Information

Table S1 reports statistics for type *B*, *C*, and *D* compounds and their pathway engagement. This material 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.

### Author Contributions

†The contributions of these authors should be considered equal.


### Notes

The authors declare no competing financial interest.

## ■ REFERENCES


(1) Peltason, L.; Bajorath, J. Systematic computational analysis of structure-activity relationships: Concepts, challenges and recent advances. *Future Med. Chem.* **2009**, *1*, 451−466.

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

(3) Esposito, E. X.; Hopfinger, A. J.; Madura, J. D. Methods for applying the quantitative structure-activity relationship paradigm. *Methods Mol. Biol.* **2004**, *275*, 131−214.

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

(5) 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.

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

(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) Yongye, A. B.; Byler, K.; Santos, R.; Martínez-Mayorga, K.; Maggiora, G. M.; Medina-Franco, J. L. Consensus models of activity landscapes with multiple chemical, conformer, and property representations. *J. Chem. Inf. Model.* **2011**, *51*, 2427−2439.

(9) 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.

(10) 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.

(11) 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.

(12) 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.

(13) 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.

(14) Tiwaria, A.; Sekhar, A. K. T. Workflow based framework for life science informatics. *Comput. Biol. Chem.* **2007**, *31*, 305−319.

(15) 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.