Evolution of the Activity Cliff Concept and Practical Implications for Compound Design

Kumulative Dissertation

zur Erlangung des Doktorgrades (Dr. rer. nat.)

der Mathematisch-Naturwissenschaftlichen Fakultät

der Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
HUABIN HU
aus Jiangxi, China

Bonn
March, 2021

Angefertigt mit Genehmigung
der Mathematisch-Naturwissenschaftliche Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Referent: Univ.-Prof. Dr. rer. nat. Jürgen Bajorath
2. Referent: Univ.-Prof. Dr. rer. nat. Michael Gütschow
Tag der Promotion: 12. Juli 2021
Erscheinungsjahr: 2021

Abstract

Entering the “big data” era, a number of different research areas have witnessed an enormous increase in data at an exponential rate. For drug discovery, various publicly available protein structure and compound bioactivity databases enable data-driven drug identification, which are further facilitated by advanced computational methods. The principle objective of structure-activity relationship (SAR) analysis is to identify structural determinants that are responsible for biological activities of compounds or other drug-relevant properties. Central to SAR analysis is the notion of molecular similarity, which can be assessed based on different principles and molecular representations. As a primary focal point of SAR analysis, activity cliffs (ACs) are receiving increased attention. By definition, ACs are formed by pairs of structurally similar compounds with large differences in potency, and thus encapsulating the notion of minor chemical modifications having large biological effects.

This thesis concentrates on large-scale AC analysis using different structural similarity and potency difference criteria, and corresponding practical implications for compound optimization in medicinal chemistry. First, AC networks, a central data structure for cliff-associated SAR analysis, was simplified yielding immediate access to SAR information. Then, a variety of molecular similarity approaches were developed, which were utilized in AC analysis to derive SAR determinants from different structural perspectives. Moreover, activity class-dependent potency difference criteria were derived by taking potency value distributions of target-based compound activity classes into account. Analyzing these similarity and potency difference criteria, this thesis represents a further evolution of the AC concept: from single- to multi-site ACs and from general to activity class-dependent AC definitions. Going beyond molecular similarity and potency difference thresholds in AC assessment, the inclusion of privileged substructures, structural isomers and single-site analogs further extended the AC concept for medicinal chemistry.

Acknowledgement

First and foremost, I would like to express my gratitude to my supervisor Prof. Dr. Jürgen Bajorath for his patience, motivation, and immense knowledge. His guidance helped me much during my research and his insightful feedbacks motivated me to sharpen my scientific thinking. I also thank Prof. Dr. Michael Gütschow for being co-referent with insightful comments and encouragement.

I further thank all of my colleagues at B-IT for unconditional support in both scientific research and daily life which made me feel like at home. I would like to thank Oliver Laufkötter and Javed Iqbal for proofreading my dissertation. Special thanks to Dr. Martin Vogt and Dr. Dagmar Stumpfe for providing many scientific suggestions, fruitful discussions to my research projects and reviewing my dissertation.

Finally, I would like to thank all of my dear friends for providing any kind of encouragements. Special thanks to my family and Hui Yu for material and spiritual support during my time in Germany.

I love all of you. Thank you so much!

Contents

Chapter 1

Introduction

1.1 Structure-Activity Relationship

Structure-activity relationship (SAR), i.e., the relation between the chemical structure of a molecule and its biological activity, encompasses the central concept that structural modifications lead to modulations of its activity.[1,2] Historically, the idea can be traced back to a publication by Crum-Brown and Fraser in 1868.[3] Deconvolution of available medical chemistry data can enable the determination of the critical chemical structural properties that are responsible for eliciting certain biological effects. For medicinal chemists, such SAR analysis has become a well-established daily routine and is frequently performed on individual, small and homogeneous compound sets which might be organized in an intuitive R-group table format for one specific target. However, when encountering much larger and heterogeneous bioactivity data, case-by-case SAR analysis is no longer feasible. For example, ChEMBL[4] (version 27) which has been manually compiled from primary published literatures, contains around two million distinct compounds with activity annotations against more than 13,000 targets, yielding a total of $\sim$16 million compound-target interactions. Such extensive heterogeneous data complicates SAR analysis, and thus efficient computational methods for comprehensive SAR studies have become a focal point of cheminformatic research. A variety of computational approaches have been introduced to perform large-scale SAR analysis such as the scaffold tree,[5] SAR matrix,[6] numerical SAR index,[7] 2D/3D activity landscape,[8,9] network-like similarity graph (NSG),[10] LASSO graph[11] or the AnalogExplorer.[12]

Beyond the identification of SAR determinants, it is equally important in SAR analysis to build mathematical models relating the chemical structure to biological activity. Derived models can be utilized for predicting the activities of untested compounds and are known as quantitative structure-activity relationship (QSAR) models.[13,14] For predictive QSAR modeling, various state-of-the-art machine learning approaches have been applied such as XGBoost or support vector regression (SVR). Apart from modeling compound activity, predictive models have been utilized for modeling other properties such as toxicity, the acid dissociation constant (pK_a), the water octanol partition coefficient (logP) or mechanisms of action for kinase inhibitors.[15–18]

A core aspect of QSAR modeling is the similarity-property principle (SPP) and its applicability forms the basis of many QSAR approaches.[19–21] The validity of the SPP can be deduced from many observations indicating that gradual structural modifications are accompanied by small to moderate potency changes, corresponding to "SAR continuity".[22] However, in violation of the SPP, pairs of structurally similar compounds with large potency differences can also occur when compounds are optimized. Especially at the hit-to-lead stage when improving compound potency is the primary task, structural neighbors of hits are heavily explored, some of which may show heterogeneous biological activity, indicating "SAR discontinuity". Large-scale data mining efforts demonstrate that SAR continuity and discontinuity often coexist in many activity classes,[22] as shown in **Figure 1**. Activity cliffs (ACs), defined as pairs of structurally similar compounds with large differences in potency, are the most prominent manifestation of SAR discontinuity.[23,24]

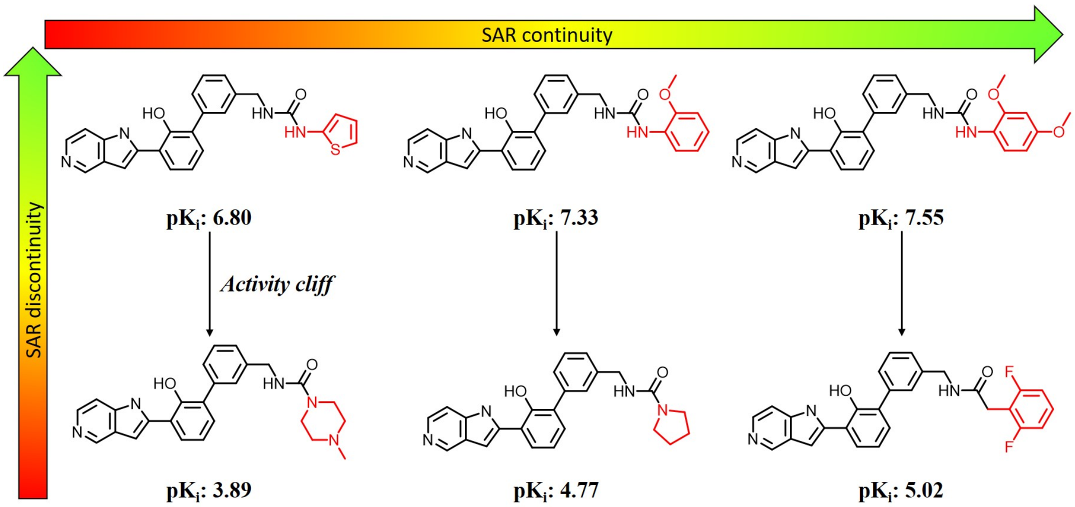


Figure 1: SAR characteristics of coagulation factor III. Shown are exemplary scenarios of SAR continuity (horizontal) and discontinuity (vertical). Compound potency values are reported (pK_i values) and structural modifications highlighted in red.

Although ACs restrain the predictive power of potency value prediction via QSAR modeling and thus are viewed controversially,[23,25,26] they also reveal the possibility of maximizing the biological response with only minimal structural modifications, and thus are highly informative in SAR analysis. In addition, AC formation is often accompanied by increased ligand efficiency, which quantifies binding affinity per atom of a ligand, and is frequently used as metric in the selection and optimization of fragments, hits, and leads.[27,28]

1.2 Molecular Representation

Typically, SAR analysis is performed based on quantitative or qualitative structural compound comparisons which strongly depend on how molecular structures are presented. In order to characterize and represent molecular structures, a variety of methods have been developed which can be roughly divided into three categories: one-dimensional (1D), two-dimensional (2D) and three-dimensional (3D) representations. Simple 1D representation includes the chemical formula of a molecule, which comprises atom types (e.g., carbon, oxygen or nitrogen) and respective atom counts. In 2D representation, the molecular structure is treated as a graph. The idea behind the molecular graph representation lies in mapping the atoms and bonds that form a molecule to sets of nodes and edges.[29] In practice, molecular structures are

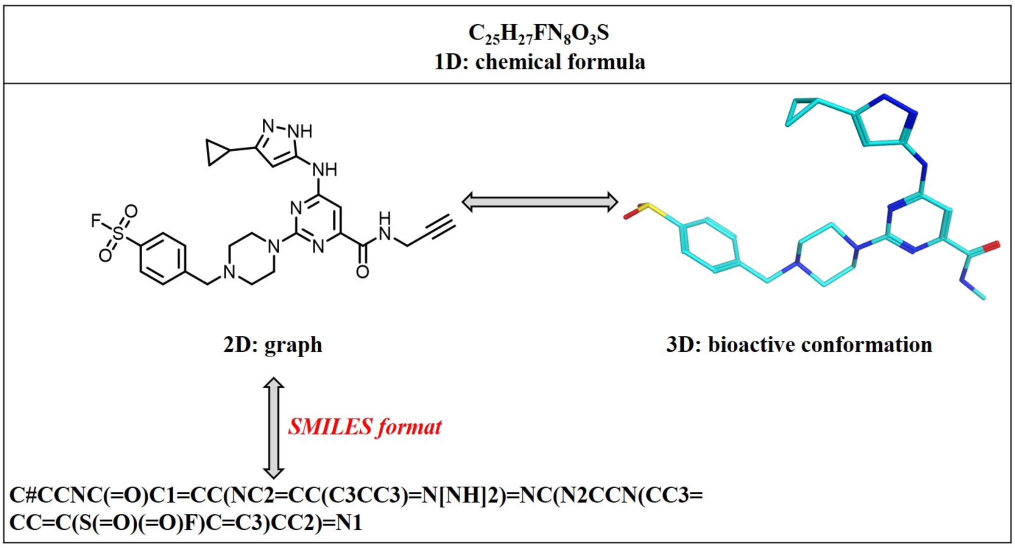

$C_{25}H_{27}FN_8O_3S$
1D: chemical formula
N-NH
NH
F
S
O
N
N
N
H
N
O
2D: graph
3D: bioactive conformation
SMILES format
C#CCNC(=O)C1=CC(NC2=CC(C3CC3)=N[NH]2)=NC(N2CCN(CC3=
CC=C(S(=O)(=O)F)C=C3)CC2)=N1

Figure 2: Molecular representations. Shown is an exemplary small molecule targeting epidermal growth factor receptor with different representations. The experimentally confirmed three-dimensional binding conformation is extracted from RCSB Protein Data Bank database with PDB code 5U8L.

frequently deposited as Simplified Molecular-Input Line-Entry System (SMILES) notation in different chemogenomic databases with the advantages of disk-efficient storage and rapid indexing.[30] SMILES is a linear and computer-friendly notation encoding the structural graph by using short ASCII strings without losing compound structure information that encompass atom types, bond types, branching, stereochemistry, cyclic, aromaticity, etc. The SMILES strings can also be easily converted into two-dimensional drawings by molecule editing software such as ChemDraw. Molecular graphs are usually depicted as planar 2D structures and, with the exception of stereochemical annotations, lack any information about the spatial arrangement of bonds and atoms. Therefore, 3D representations are introduced which contain such information by taking conformational flexibility of a molecule into account.[31] Hence, while 1D and 2D representation are well defined by the chemical compound, multiple (reasonable) 3D representations of a compound can exist depending on the its conformation(s). Exemplary molecular representations are depicted in **Figure 2**

1.3 Molecular Fingerprint

Given a molecular representation, the structural characteristics of small molecules need to be converted into a computer-readable form in a way that allows the comparison between molecules. Most commonly, continuous or binary numerical values known as molecular descriptors are determined from molecular representations resulting in multidimensional feature vectors that represent certain molecular properties of interests (e.g., substructures, physiochemical properties, topology, or pharmacophores).[32] Two-dimensional (2D) fingerprints derived from molecular graphs, such as the molecular access system (MACCS)[33] and extended-connectivity fingerprint (ECFP),[34] are some of the most popular descriptors in characterizing chemical structures. In the MACCS fingerprint, each predefined chemical substructure is represented by a fixed vector position with a length of 166 bits (i.e., 166 structural patterns), each of which accounts for the presence or absence of a specific structural pattern (**Figure 3a**). ECFP fingerprints are circular topological fingerprints, which capture local atom environments by considering circular layers centered at each non-hydrogen atom with increasing bond diameter up to a predefined maximum. ECFP4 encodes three layers of circular atom environments up to a maximum bond diameter of four (**Figure 3b**). Applying a hashing function, patterns of

ECFP fingerprints can be encoded into a single integer value and further folded into a fixed length fingerprint of commonly used 1024- or 2048-bits.

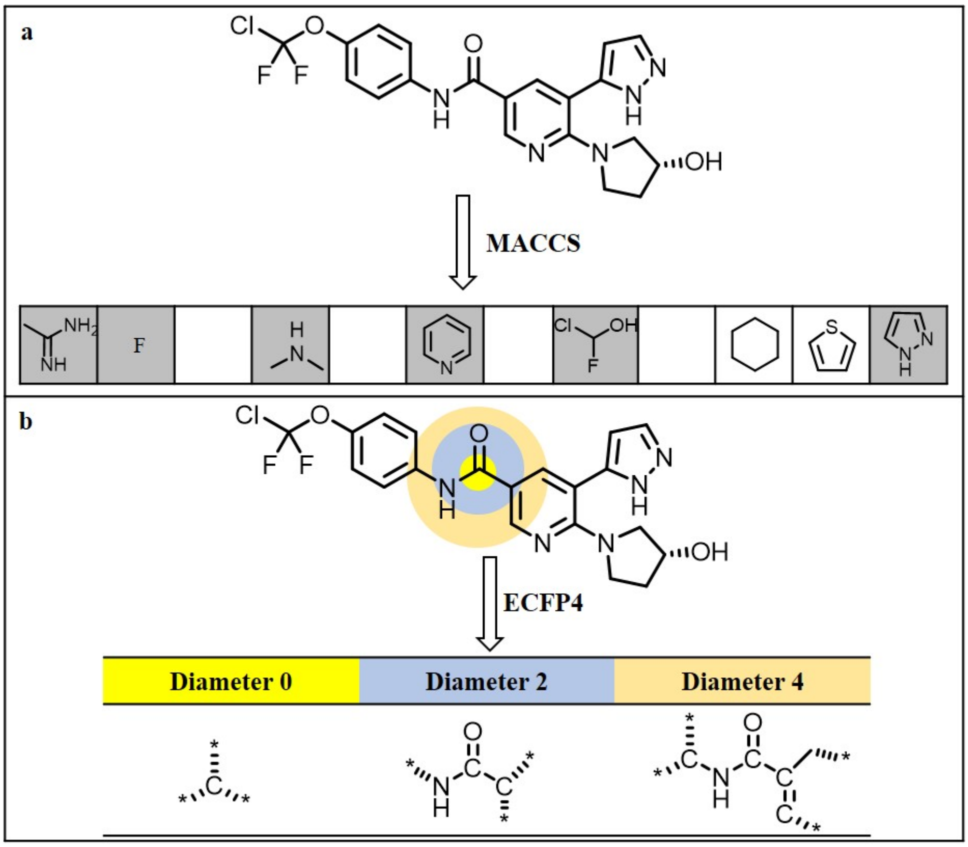


Figure 3: Molecular fingerprints. Shown are a typical substructure-based fingerprint (a) and a circular fingerprint (b) used for characterizing a molecular structure. In (a), the MACCS fingerprint is represented by 166 predefined substructures placed at fixed vector positions. In the vector, the corresponding bit is set on (grey background), if the predefined structure pattern is presented in the molecule; otherwise, the bit is set off indicating the absence of particular features. In (b), the calculation of an ECFP fingerprint with bond diameter four centered on a carbon atom (yellow) is illustrated. The resulting topological environments with increasing diameter representation are shown. The connectivity information beyond the observed diameter is given by dashed lines, with dummy atoms represented by asterisks (*).

1.4 Molecular Similarity

How to assess and quantify molecular similarity represents one of the most fundamental and intensely studied topics in cheminformatics.[2,31,35] Molecular similarity assessment might seem a deceptively simple question for a specific compound optimization project. However, in order to systematically evaluate and extract structurally similar compounds from large volumes of compound depositories covering an extensive chemical space, molecular similarity assessment must be clearly stated out and consistently applied. A variety of computational approaches for molecular similarity assessment have been established from quantitative or qualitative perspective yielding continuous numerical or binary values, respectively.

1.4.1 Numerical Similarity Metrics

1.4.1.1 Fingerprint-Based Similarity

In cheminformatics, fingerprint-based methods are often applied for compound similarity assessment. Molecular similarity can be quantified using various similarity indices usually ranging from 0 (no overlapping fingerprints) to 1 (identical fingerprints).[31] The most popular index is the Jaccard index or Tanimoto coefficient (Tc), which is defined by

$$\mathrm{Tc}\left(\mathrm{A},\mathrm{B}\right) = \frac{c}{a+b-c}$$

Here, a and b are the number of chemical features detected in compound A and B, respectively, and c denotes the number of chemical features shared by both compounds.[36] Given that Tanimoto-based similarity accounts for whole-molecule similarity, a high Tc value sometimes does not necessarily correlate with a high structural similarity of compound pairs from a chemical viewpoint. Even If two compounds have a Tc value of 1, it does not indicate two compounds are identical since this formula does not consider chemical feature frequency or how detected structural patterns are connected to each other. Similarity of compounds is often determined by setting a threshold to judge whether compound pairs should be viewed as sufficiently structurally similar. However, there is no universally applicable threshold since the calculated Tc values are fingerprint-dependent. In general, Tc values of 0.55 and 0.85 are widely used for ECFP4 and MACCS fingerprints, respectively.[37] At the top of **Figure 4**, an exemplary compound similarity calculation on the basis of MACCS fingerprint is shown.

1.4.1.2 3D-Based Similarity

Three-dimensional molecular representation allows compounds to be compared based on the spatial arrangements of atoms and bonds. For ligands, this requires knowledge of 3D conformations that can be obtained either using experimental methods such as X-ray diffraction techniques or by computational methods. On the basis of protein-ligand crystal structures stored in the PDB database,[38] three-dimensional

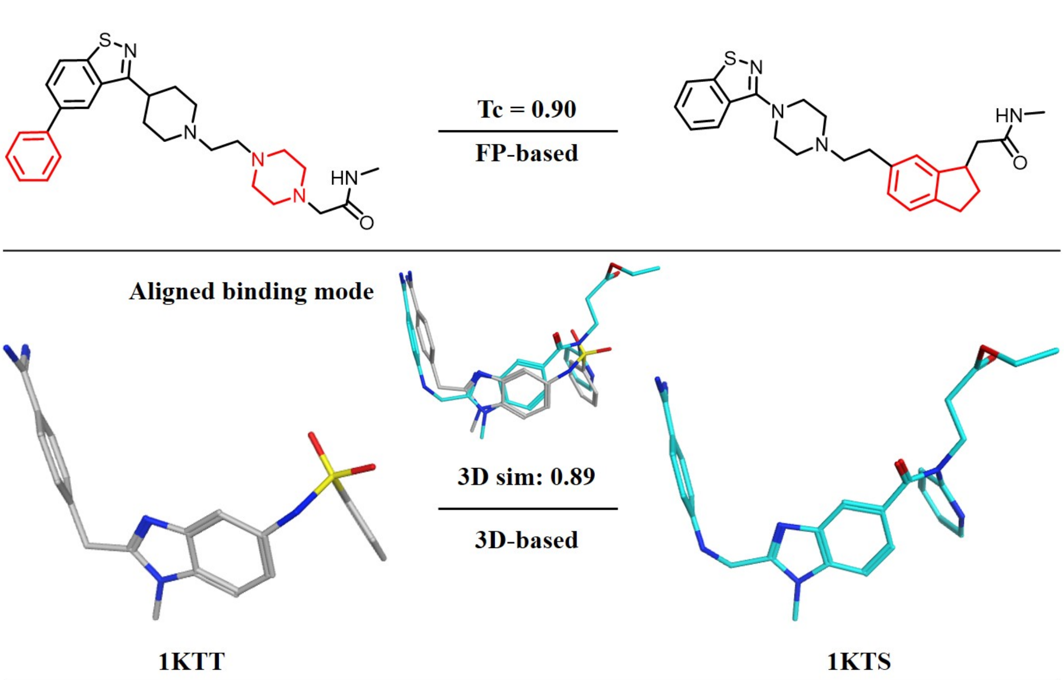


Figure 4: Predefined thresholds to assess compound similarity. Depicted are exemplary approaches applying fingerprint- and 3D-based methods to define structurally similar compound pairs. For the fingerprint approach, the MACCS fingerprint is used. For three-dimensional assessment, the binding modes of two ligands co-crystalized with prothrombin are superposed and aligned to yield a 3D similarity value. Structural modifications are highlighted in red in the structural graph representations.

binding modes of ligands can be isolated from the complexes. To systematically quantify the 3D similarity of paired ligands, a computational method has been introduced to explore 3D similarity relationships between a set of crystallographic ligands.[39,40] This methodology is based on the use of property density functions taking the binding conformations and orientations (3D coordinates) of paired ligands into account. Accordingly, the first step is the superposition of α-carbon atoms of selected protein structures that can be used for comparing the positional difference of paired ligands. Then, a property density function is computed for each ligand accounting for four different atomic properties (aromaticity, hydrophobic, hydrogen-bond acceptor and donor characters) and the atomic coordinates. The 3D similarity values are calculated according to the normalized overlapping density function of paired ligands. Exemplary compound pair with high 3D similarity value is illustrated in **Figure 4** (bottom). In the event that no experimentally confirmed binding modes are available, putative 3D confirmations can be generated by computational algorithms, such as the OMEGA toolkit implemented in OpenEye scientific software.

Then, a shape-based superposition method (ROCS) is applied which quantified the molecular similarity on the basis of shape-based similarity metrics. ROCS aligns two molecules using a solid-body optimization process that maximizes the overlap volume between them.[41,42] The 3D similarity evaluations also require a predefined similarity threshold. However, 3D similarity evaluations provide a new angle for analyzing SAR information and are often much more intuitive and practical for analyzing interactions, or deriving potential target annotations for database compounds with complex structures (e.g., natural products).[41]

1.4.2 Substructure-Based Similarity

Substructure-based similarity measurement derived from 2D graph representations is a popular alternative similarity assessment method and, due to its interpretable and chemically intuitive nature, largely appreciated by medicinal chemists. In this approach, the existence of a common core structure of significant size is a prerequisite for considering two molecules similar or not, yielding a binary similarity assessment. Hence, it circumvents the abstract nature of similarity calculations based on numerical similarity assessments, which require predefined similarity thresholds. Different substructure-based methods for determining the common core structure have been suggested such as the formation of matched molecular pairs[43] or analog series-based scaffolds,[44] as reported herein.

1.4.2.1 Matched Molecular Pairs

By definition, a matched molecular pair (MMP), also termed single-site analog pair, is a pair of compounds that are only distinguished by a chemical modification or an R-group replacement at a single site.[43] Accordingly, an MMP can be characterized by a shared core and two R-groups (i.e., a "chemical transformation"),[45,46] as shown in **Figure 5a**. Because of its chemically intuitive and easily interpretable nature, the MMP concept is appealing for medicinal chemists and has gained wide popularity in the scientific community in recent years. From a computational aspect, such pairs can be efficiently and easily identified in large compound databases through the application of compound fragmentation algorithms that systematically generate all feasible cores and R-groups of a compound.[46] In order to generate R-groups that are typically observed in medicinal chemistry, additional criteria can be applied during MMP generation. So called size-restricted MMPs[45] are generated using the following criteria: (i) the heavy-atom count of the core structure must be at least twice as large as that of the R-group; (ii) R-group must not contain more than 13 heavy atoms; (iii) the size difference between two exchanged R-groups must not exceed eight heavy atoms.

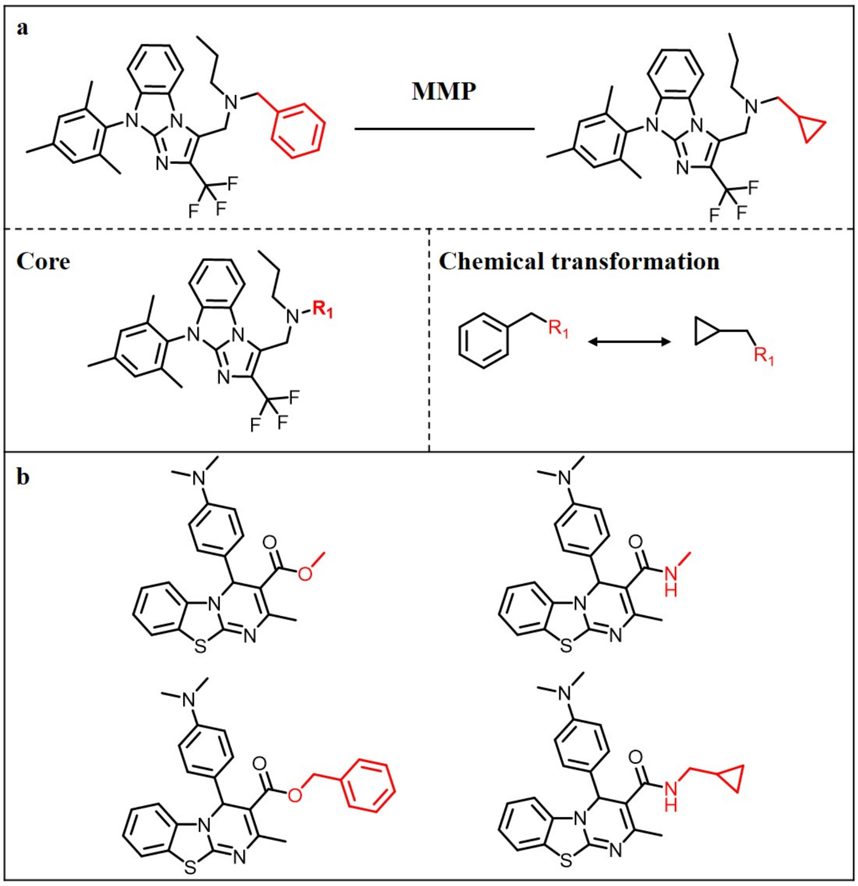


Figure 5: Single-site analogs. Shown are exemplary analogs forming a pair (a) or a series (b). For the MMP relationship (a), the core structure and chemical transformation are provided. The structural modifications (exchanged substituents) are highlighted in red.

By default, MMPs are generated considering arbitrary acyclic single bonds. By requiring fragmentation bonds to follow predefined retrosynthetic rules (RECAP rules),[47] so-called RECAP-MMPs or RMMPs are generated, which take synthetic accessibility into account.[48] As an extension of the MMP concept, matched molecular series (MMS) is generated which is defined as a series of two or more analogs (≥ two compounds) with chemical modifications at a single site (**Figure 5b**).[49,50] The utilization of the MMP or MMS concept facilitates the analysis of structure-activity or structure-kinetic relationships (e.g., on- and off-rates of binding),[51] or provides the rationale suggestion of suitable R-groups to be explored in compound optimization stages.[49]

1.4.2.2 Analog Series

An analog series (AS) can be defined as a set of compounds that share the same core structure but carry different R-groups at the same or different core substitution site(s). Previously, two computational approaches have been introduced to systematically extract analog series from large compound data sets. The first one, based on the identification of RMMP clusters, was introduced five years ago.[44] Here, for a given data set, compounds are first collected and subjected to bond fragmentation to generate RMMPs. Then, all RMMPs are organized into a network where nodes indicate RMMP-forming compounds and edges pairwise RMMP relationships. In the network, each disjoint connected component (cluster) is viewed as an individual AS.

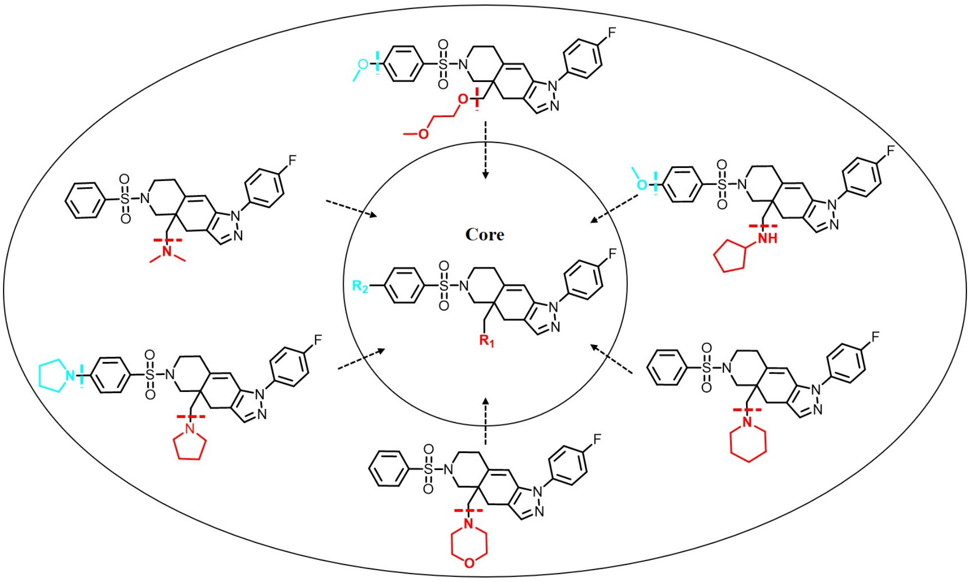


Figure 6: Illustration of compound-core relationship method. Shown is an analog series (AS) generated by the CCR algorithm. Six analogs are fragmented according to retrosynthetic reaction rules as indicated by the dashed colored lines. After compound fragmentation, the compounds sharing the same core are grouped together and viewed as analogs. Here, the analogs are represented by an AS with two substitution sites (R_1 and R_2) highlighted in red and cyan, respectively.

Recently, another methodology on the basis of compound-core relationships (CCRs)[52] was proposed to systematically organize compounds into different analog series (**Figure 6**). The CCR method is mainly comprised of three steps. Firstly, all compounds are subjected to bond decomposition according to RECAP rules, permitting at most five substitution sites per compound. Then, similar to MMP generation,[45] size restrictions are enforced to confine R-group sizes to at most 13 heavy atoms. Lastly, upon bond fragmentations, the cleavage sites of the cores are replaced by hydrogen atoms, yielding so-called generalized cores and all compounds sharing the same generalized core are organized into individual analog series. As a result, by adding appropriate substitution sites to the generalized cores, the compounds in one AS can be easily visualized in a R-group table format.

1.4.2.3 Scaffolds and Privileged Substructures

Scaffolds are predominately used to describe the core structures of a set of compounds.[53,54] Core structures can be defined in a variety of ways and are subjective by nature. In one specific medicinal chemistry project, especially in late stage lead optimization, fine-tuning of interesting molecular structures is often performed in order to improve pharmacodynamic properties. In this scenario, the core structure is easily captured. However, systematically extracting core structures from chemical libraries require a predefined core structure definition and time-efficient computational algorithms.[55] A first formal and widely used definition of scaffold is the Bemis-Murcko scaffold (BM scaffold).[56] The essence behind BM scaffolds is that compounds typically comprise three components, i.e., ring systems, chemical linkers between rings, and R-groups attached at rings and/or linkers. If the branches (R-groups) are removed, the remaining parts, i.e., ring system and linkers, are termed the BM scaffold. BM scaffolds can be further simplified by converting all non-carbon heavy atoms and non-single bonds into carbon atoms and single bonds, respectively, yielding a more generic cyclic skeleton (CSK),[57] as shown in **Figure 7**.

Although BM scaffolds could be used to organize a set of compounds, this approach has certain limitations. Since ring systems are central in defining BM scaffolds, adding or deleting one ring, or just converting a single heavy atom to another one in a ring (e.g., C $\longleftrightarrow$ N) will easily yield a new BM scaffold. As such, BM scaffolds are not consistent with analog generation through chemical reactions where simple five- or six-member rings could also be viewed as attached R-groups to the core.[58]

To overcome this shortcoming, a new scaffold derived from analog series (analog series-based scaffold or ASB scaffold) was introduced.[59,60] Following the RMMP-based network approach as mentioned above, one disjoint cluster represents an individual analog series comprising at least one RMMP. In the network, all possible RMMP cores are enumerated in a specific cluster. If analogs in one cluster are only distinguished by a single substitution site, the minimal RMMP core structure (i.e., the core with the minimum number of heavy atoms) is taken as the scaffold of this

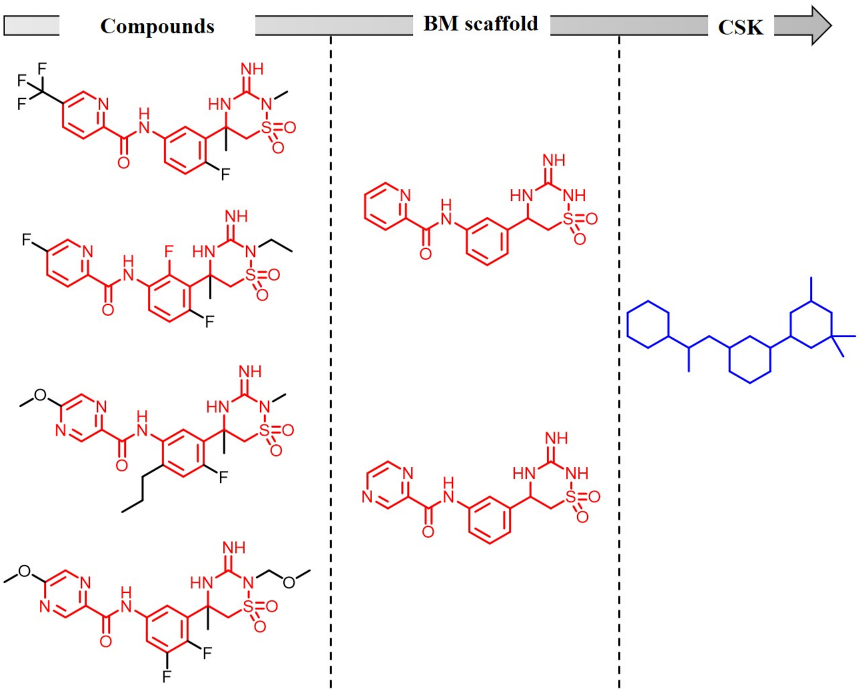


Figure 7: Compound-scaffold-CSK hierarchy. The stepwise generation of BM scaffolds and CSK skeleton from four exemplary small molecules is depicted. The two BM scaffolds colored red are topologically equivalent, yielding a single common CSK skeleton (blue)

AS, representing the first generation ASB scaffold which covered around 70% of ASs in ChEMBL 22 database.[44] The ASB scaffold definition was further extended by considering the ASs with multiple substitution sites. For such ASs, the overlapping part of all RMMP cores leads to the introduction of second generation ASB scaffolds with a coverage of more than 90% of ASs in the ChEMBL database.[60,61] To some extent, this network-based method can avoid the limitations of BM scaffolds, it takes retrosynthetic information into account and is not confined to a restricted molecular hierarchy. However, for some extremely complex clusters representing around 10% of ASs in ChEMBL 22 database, it was still difficult to extract one unambiguous ASB scaffold to represent the corresponding AS.[60,61] The recently developed CCR method (as described above) has been a significant and robust improvement to address these issues. It is applicable to large data sets and able to identify large AS with single unique core structures containing multiple substitution sites.[52]

The concept of core structures and the detection of ASs make it possible to identify core structures shared by many compounds that are enriched within specific data sets. When biological activities are associated with each compound, scaffold identification for ASs allows the exploration of “structural motifs” or “privileged substructures” having specific target family preference.[62] The idea of privileged substructures was firstly put forth by Evans in the late 1980s when he was studying the benzodiazepine nucleus.[63] Given the appeal of this notion, many so-called privileged substructures were identified, typically through frequency of occurrence analysis of chemical entities, in different therapeutic target families. The most comprehensive privileged substructure compendium extracted from drugs and natural products was provided by Welsch *et al.*[64] These frequently occurring building blocks indicate their usefulness for designing drug-like compounds targeting a desired target family but do not necessarily imply they are selective for that family. Indeed, firm evidence indicates that “target family-directed privileged substructures” might not truly exist. Instead they are frequently detected in completely different target families[65] such as the benzodiazepine scaffold found in many ligands of G protein-coupled receptors (GPCRs), ion channels and protein kinases; similarly, compounds containing biphenyl[66] or the indole moiety,[67,68] can be found in compounds active against distinct targets (**Figure 8**). However, due to the drug-like properties and high tendency to preferentially bind to specific target families, these PSs will continue to be of high interest as a starting point to design novel bioactive compounds.

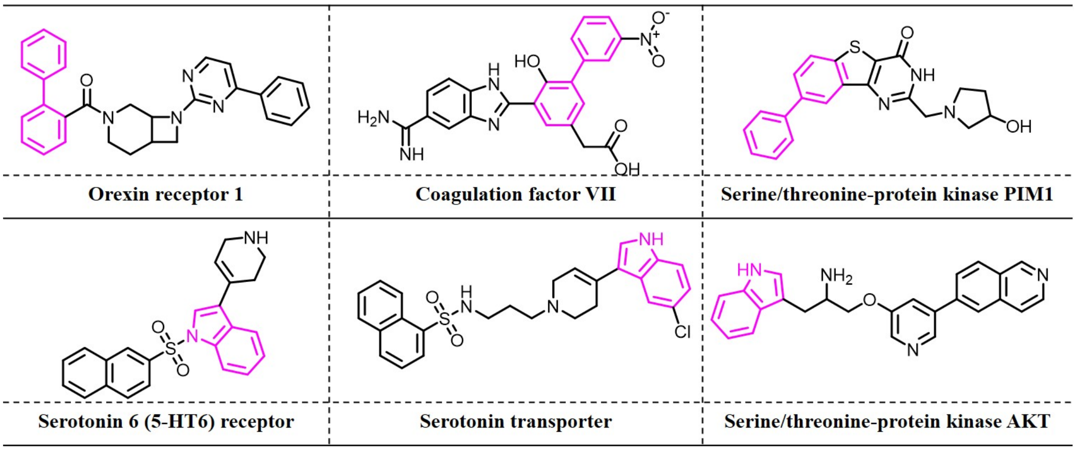

Orexin receptor 1
Coagulation factor VII
Serine/threonine-protein kinase PIM1
Serotonin 6 (5-HT6) receptor
Serotonin transporter
Serine/threonine-protein kinase AKT

Figure 8: privileged substructure-containing bioactive compounds. Shown are two exemplary scaffolds, i.e., biphenyl and indole, embedded in different molecules with diverse target annotations. The privileged substructures are colored pink and the target annotations (below) are provided.

1.5 Progress in Activity Cliff Research

1.5.1 Activity Cliff Criteria and Categorization

ACs represent the most prominent instances of SAR discontinuity and reveal SAR determinants, and thus are highly informative in SAR analysis. ACs are defined as pairs of structurally similar compounds with a large difference in potency. For a rigorous definition of ACs, two criteria must be clearly specified: (i) when should a pair of compounds be considered structurally similar, (ii) what is the minimum potency difference required to indicate SAR discontinuity for a specific target. Different evaluations in assessing molecular similarity have been proposed, such as the comparison of fingerprints or the presence of common core structure, as mentioned above.[24] To ensure the reliability of AC assessment, the use of high quality activity annotations with assay-independent equilibrium constants (K_i values) is highly recommended.[69] Unlike structural similarity assessment, little attention has been paid to investigating to when a potency difference might be considered statistically significant and sufficiently large for AC formation. Instead, a 100-fold change in potency is frequently applied irrespective of the activity classes.[24] The combination of different similarity assessments and an at least 100-fold difference in potency leads to the introduction of different AC categories. Different generations of ACs allow SAR determinant explorations from diverse structural perspectives, and thus mirrors the evolution of the AC concept.[24,70]

1. Fingerprint-based activity cliff: high Tanimoto coefficient (Tc) values according to fingerprint comparisons ($Tc \geq 0.85$ in the case of MACCS fingerprint) plus at least two orders of magnitude (100-fold) difference in potency.[71]

2. Substructure-based activity cliff: the presence of a common core structure for paired compounds such as the formation of MMP relationships or compounds differing by the configurations at a single stereo center with at least two orders of magnitude difference in potency.[45]

3. Three-dimensional activity cliff (3D-cliff): compounds show at least 80% similarity of experimentally confirmed binding modes but have at least two orders of magnitude difference in potency.[40]

Exemplary ACs are shown in **Figure 9**. Specifically, ACs based on chemically intuitive similarity assessments, like MMP-cliffs or chirality cliffs, have been intensely studied[24,45,72] and various machine learning models have been developed to differentiate cliff and non-cliff pairs.[73–76]

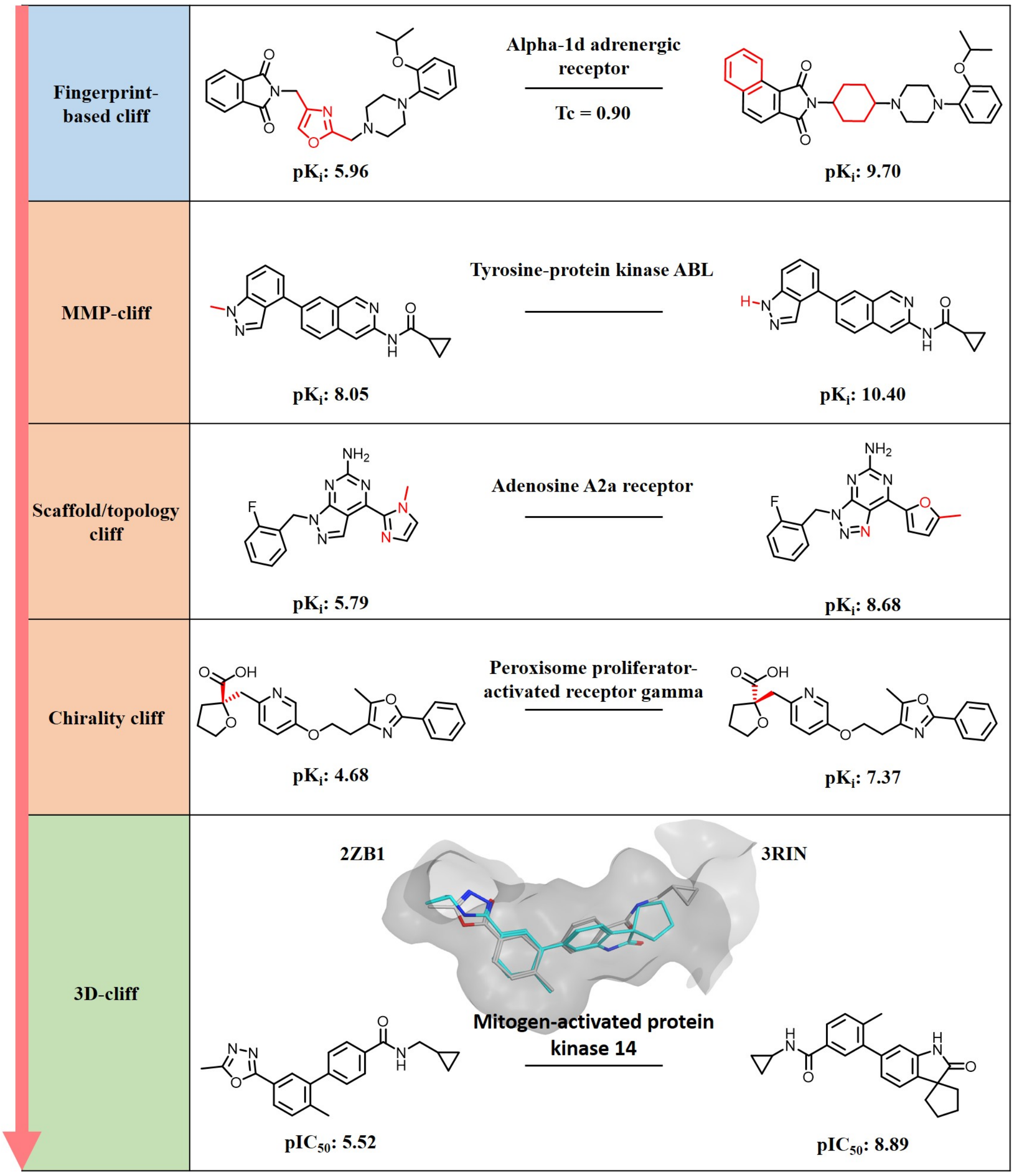


Figure 9: Activity cliff categorization. From top to bottom, shown are five exemplary ACs evaluated by comparing MACCS fingerprints (blue), common substructures (orange) and binding modes (green). For each AC type, the structural modifications are highlighted in red. Scaffold/topology cliff indicates that pairs of compounds share topologically equivalent scaffolds but have differing R-group topologies.[72] The corresponding target annotations and compound potency (reported as pK$_i$ or pIC$_{50}$ values) are given. In addition, the aligned binding mode for the compounds forming an 3D-cliff is provided and PDB codes are reported.

1.5.2 Inactive Compounds in Cliff Analysis

High-throughput screening (HTS) provides an opportunity to identify novel hits by automatically assaying large compound depositories against variable targets.[77] However, even in large-scale screening projects, the hit rate from experimental HTS is often low, which in turn yields a large fraction of inactive chemical entities.[77–79] According to the screening data deposited in the PubChem BioAssay database, a surprising proportion of small molecules tested in ≥ 100 assays are consistently inactive, an observation which is also referred to as dark chemical matter (DCM).[80,81] The screening results do not necessarily indicate inactive compounds are biologically inert. Instead, a lot of studies demonstrated the attractiveness of DCM.[80] For example, Ballante *et al.* identified GPCR ligands with sub-micromolar affinities with the aid of molecular docking to screen a commercially available fraction of DCMs against A_{2A} adenosine and D_4 dopamine receptors.[82] Hence, if appropriate targets for these neglected subsets are identified, they could be highly selective.

Apart from target identification for inactive molecules, additional knowledge can be gained by involving confirmed inactive compounds in SAR studies. Attempts have been made to systematically extract ACs formed by confirmed active and inactive molecules from the PubChem database.[83] For defining ACs involving inactive compounds, the potency difference threshold is not applicable. Instead, in order to avoid weakly potent compounds forming ACs with inactive compounds, a potency threshold is applied to define active compounds; specifically, a minimum potency value of 10 μM is often required.[83] ACs involving inactive compounds provide additional insights in understanding how small chemical modifications could transform (highly) potent active compounds to compounds with essentially no biological response. Such information is extremely valuable at the early stage of SAR exploration, especially for underexplored targets with only a limited number of known bioactive compounds.

1.5.3 Activity Cliff Characteristics and Visualization

A large-scale study of ACs in the ChEMBL database revealed that around 5% of structurally similar compound pairs met AC potency difference criteria which were formed by $\sim$25% of all available high-confidence bioactive compounds.[84,85] Most of the identified ACs were formed between compounds within the micromolar and nanomolar potency range.[84]

No. clusters (clusters ≥ two ACs)	No. ACs (coordinated ACs)
80 (53)	810 (783)

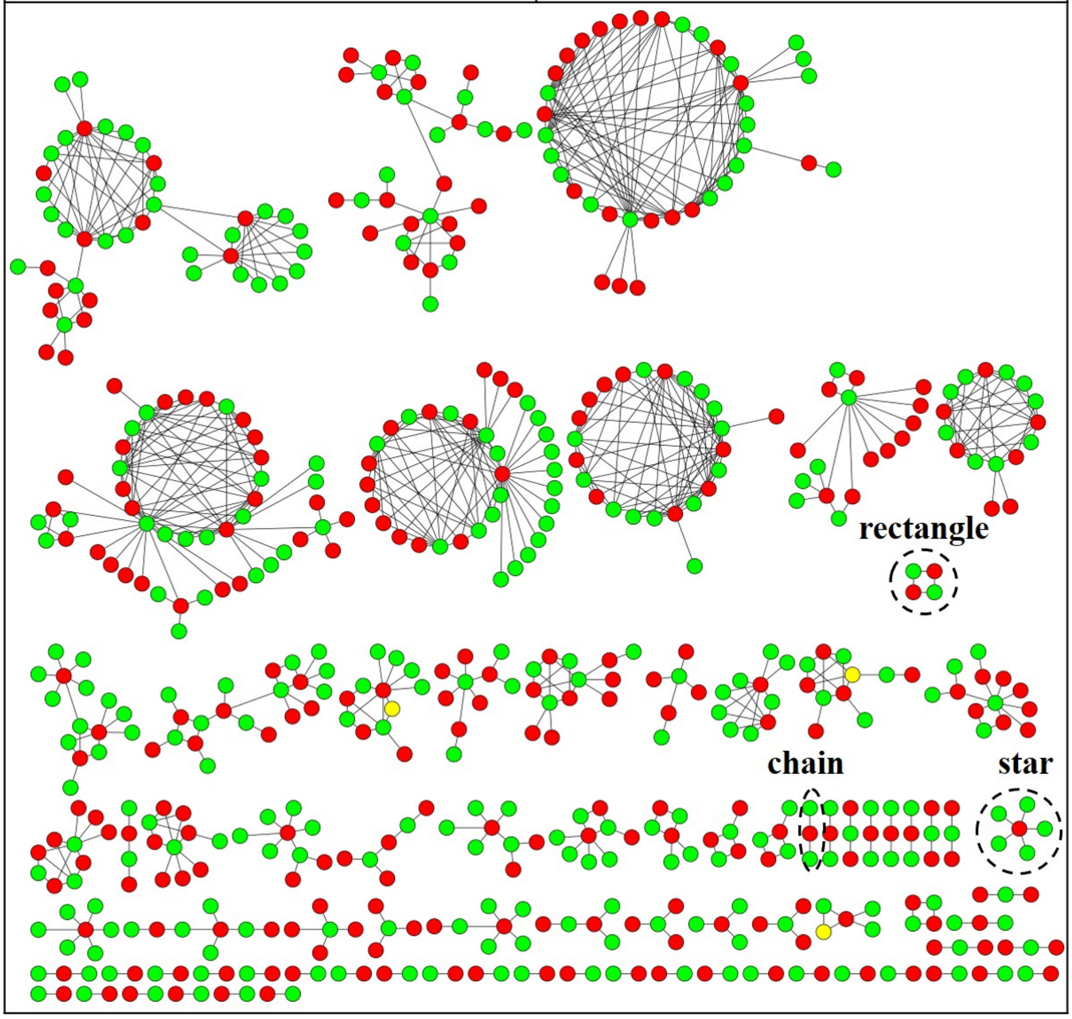


Figure 10: Exemplary activity cliff network. Shown is an exemplary MMP-cliff network formed by cannabinoid CB2 receptor ligands. In the network, nodes represent AC compounds and edges pairwise MMP-cliff relationships. In addition, nodes are colored green, red or yellow if AC compounds are highly, weakly or highly/weakly potent cliff partners, respectively. Representative recurrent topologies are encircled with black dashed lines. Statistical network analysis is provided.

Although ACs are defined on the basis of compound pairs, ACs are not typically formed in an isolated manner. On the contrary, more than 90% of them are formed in a coordinated manner where compounds are involved in more than one AC, a phenomenon that can be observed irrespective of the molecular representations. This is shown in AC networks where nodes represent AC-forming compounds and edges pairwise AC relationships (**Figure 10**).[71] In AC network representations, the disjoint AC clusters often showed different AC compositions and network topologies such as recurrent stars, chains and rectangles within these clusters.[86] The coordinated manner of ACs leads to the introduction of the activity ridge concept, which comprises multiple highly and weakly potent compounds where each possible pair of highly and weakly potent compound forms an AC.[87] In addition, AC compounds having an extremely high propensity to form ACs (i.e., "hubs" in densely connected AC clusters) are termed "AC generators".[8] Taken together, coordinated ACs giving rise to varying size of clusters are much more informative than isolated ACs. Case-by-case SAR analysis of sub-clusters requires interactive network analysis, however, how to systematically extract and analyze SAR information from clusters especially these densely connected clusters requires advanced computational methods.

1.5.4 Cliff-Associated SAR Analysis

Large-scale AC analysis or AC prediction falls into the domain of cheminformatics. ACs encapsulate important SAR information, although, how to appropriately interpret and communicate this information to medicinal chemists for guiding compound optimization is a non-trivial task.[24] Indeed, for more than 75% of MMP-cliffs, no evidence was observed indicating that advanced analogs of highly potent cliff-forming compounds had been identified. The limited utilization of AC-based SAR information might be attributed to AC data not being immediately accessible to medicinal chemistry in a chemically intuitive format. On the other hand, for around 25% of ACs, further chemical modifications of highly potent cliff partners have been observed, among which for $\sim$15% of ACs more potent analogs were identified indicating optimization efforts.[88]

Given the underutilization of the AC concept and complexity of some AC clusters, attempts have been made to systematically extract SAR information from AC clusters by applying computational methods. Taking MMP-cliffs as an example, numerical cluster indices, i.e., the MMP index and MMP core index, were introduced to correlate structural similarity of AC-forming compounds and AC diversity. The MMP index value indicates the proportion between the number of existing MMPs and the number of theoretically possible MMPs formed by compounds within a given cluster which quantifies the degree of structural similarity across AC compounds. Whereas, the MMP core index is the ratio of the number of cliff cores relative to the number of MMP-cliffs which indicates AC diversity. This methodology was designed to characterize and prioritize AC clusters, and systemically organize these clusters into an index map which was further divided into four regions according to both index values.[89–91] Hence, this approach provides partial cliff-associated SAR analysis by reorganizing AC clusters according to numerical index values indicating the complexity of SAR information within the clusters.

Alternatively, compounds in one disjoint AC cluster might be distinguished by single and/or multiple substitution site(s). If an AC cluster contains analogs differing only at a single substitution site, by definition, it can be organized into one MMS which provides a simple scenario for SAR analysis. If an AC cluster contains analogs differing at multiple substitution sites, in principle, it contains at least two MMSs. MMSs can be paired by identifying compounds shared between two series. Accordingly, SAR information can be viewed in a pairwise manner based on MMSs.[50] Paired MMSs sometimes exhibit the transfer of SAR information from one series to another series.[92] Thus, this approach provides another angle for cliff-associated SAR analysis by dissecting AC clusters according to shared AC compounds.

1.5.5 Activity Cliff Rationalization

Since ACs indicate minor chemical modifications with large potency effects, the induced effects on binding may imply critical protein-ligand interactions or binding conformational changes which could be explained with the aid of crystallographic structures, if available.[93] In the cases of 3D-cliffs, which take the experimentally confirmed ligand binding modes into consideration, clear interaction differences for $\sim$40% of 3D-cliffs could be attributed to lipophilic/aromatic group-associated hydration or shape complementarity effects, followed by H-bond and/or ionic interactions accounting for around 30% (**Figure 11**). For a small fraction of AC instances ($\sim$0.4%), the interaction differences due to the presence of water-mediated hydrogen bonds were detected. These structural waters involved in specific interactions are often considered as a part of the protein structure.[40,94,95]

A more systematic and accessible method for the detection of the interaction hotspots or differences depending on generation terms protein-ligand interaction fingerprints. This fingerprint format encodes interactions as bit vectors with "1" and "0" indicating the presence or absence of specific interactions, respectively.[96] Interaction fingerprints have been widely used in two major therapeutic target groups, protein kinases and GPCRs.[97–99] In analogy to fingerprint-based ACs, interaction cliffs were introduced and defined as pairs of ligands extracted from ligand-target complexes with high structural and interaction similarity and a large difference in potency.[100] Interaction fingerprints provide another angle for rationalizing AC for-

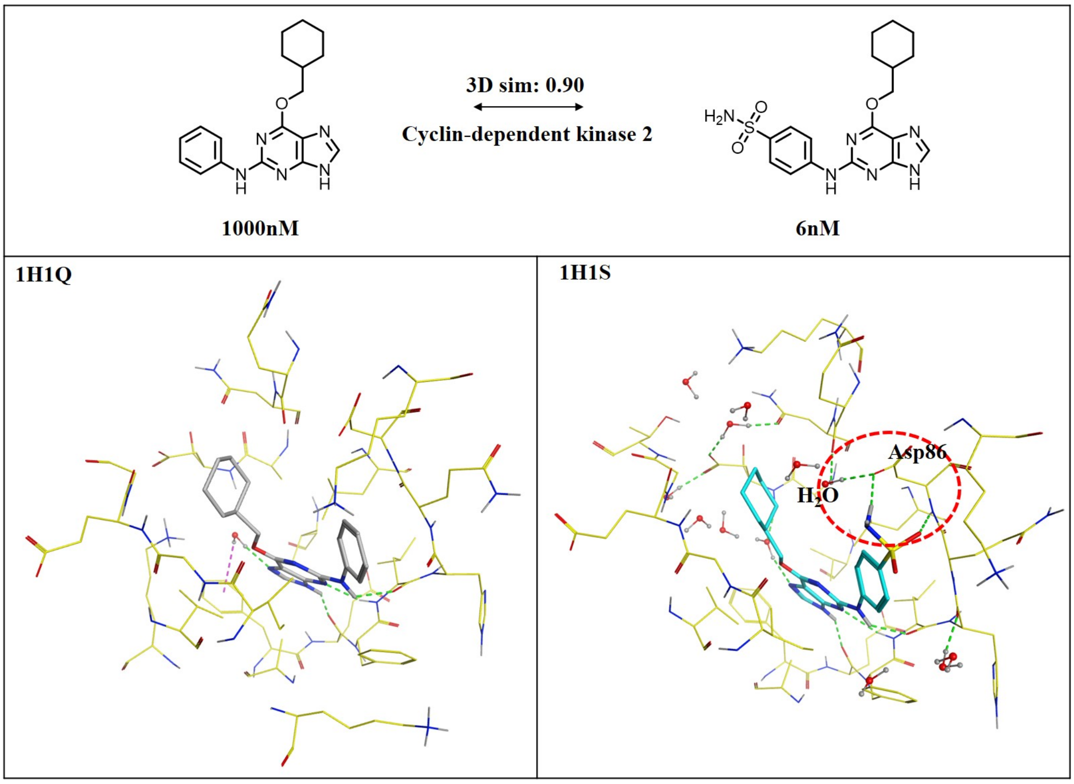


Figure 11: Exemplary three-dimensional activity cliff. Shown are an exemplary 3D-AC extracted from cyclin-dependent kinase 2 target and the corresponding binding mode analysis. The carbon atoms of the protein and the highly and weakly potent cliff ligands are colored yellow, cyan and grey, respectively. The interaction differences are highlighted using red dashed circle. PDB IDs of X-ray complexes (upper left), compound structure and potency values are provided. In addition, the green dashed line indicates hydrogen bond formation.

mation and interaction cliffs can help to detect interaction hotspots. However, this approach might fail to prioritize key interactions which contribute more to the total free binding energy.[101]

Taken together, these detectable interaction differences between highly/weakly potent cliff-forming compounds can aid in the identification of the interactions between key residues and ligand atoms that are important for molecular recognition. Hence they provide structural rationale for activity cliff formation which is beneficial to structure-based drug discovery.[93,94] However, a certain percentage of ACs remain difficult to interpret and rationalize such as the recently introduced off-pocket ACs[102] whose modified sites are located in solvent-exposed areas. The absence of obvious interactions between ligand atoms and residues complicates the AC rationalization, and thus hinders their practical applications in medicinal chemistry. Even for these identified interaction differences, further experimental tests are needed to confirm the binding free energy contribution of specific interactions. In the instance of the "magical methyl" group, the introduction of such a group yielding $\geq$ 100-fold boost in potency stems possibly from the cooperative interplay of conformational, hydrophobic, desolvation or other effects, rather than being attributed to one specific interaction.[103] Moreover, the target-ligand complex provided by X-ray diffraction is a static and time-average view under experimental crystallization conditions. It does not reveal dynamics of protein-ligand interactions, which include multiple aspects such as desolvation, entropic penalty due to configurational restriction, enthalpic effects, and other long-distance interactions.[95,104–106] Additionally, lack of paired X-ray complexes for cliff-forming ligands (as one might expect, databases are typically devoid of the structural complexes of the weakly potent cliff compounds and their biological targets), further complicates the AC analysis on the basis of structural data. In these cases, more advanced computational methods for dynamic molecular modeling such as molecules dynamics (MDs) are needed to probe the possible mechanistic reasons for molecular recognitions between structurally similar cliff molecules.[107,108]

1.6 Thesis Outline

This PhD thesis focuses on the further investigation of ACs. Different computational methods evaluating compound similarity are developed, mirroring the continued evolution of the AC concept. Furthermore, the practical implications of ACs in compound design are discussed.

- Given the popularity of network representation in AC analysis, in *chapter 2* a methodology aiming to reduce network complexity and to easily access the SAR information stored in a network is reported. This approach is built upon the MMS concept and pairing series according to second-round fragmentation. Simplified networks could serve as complementary and easily accessible structures to visualize and analyze cliff-associated SAR information within the original AC network.

- *Chapter 3* extends the MMP-cliff data structure through the introduction of structural isomers for MMP-cliff compounds. The combination of structural isomer and matched molecular pair relationships yields a new AC category, which encodes the potency effects of the moving substituents around different core positions.

- *Chapter 4* systematically explores the frequency of occurrence of privileged substructures (PSs) in bioactive compounds deposited in the ChEMBL database. Different molecular properties are used to differentiate the ACs with PSs and without PSs. In addition, PS-based AC networks are constructed to analyze structure context-dependent biological activities.

- *Chapter 5* reports a unified strategy for extracting different types of graph-based ACs reported in former chapters. For ACs differing at two substitution sites, a four-compound data structure for dual-site ACs is suggested for SAR analysis. All identified ACs have been made available to the public for follow-up analysis.

- *Chapter 6* systematically extracts ACs that capture minimal structural alterations, i.e., heteroatom walk or replacement. For these 2D-cliffs, search for X-ray complexes of cliff compounds and their cliff targets is conducted in the PDB database to rationalize AC formation at atomic levels.

- *Chapter 7* explores the difference of AC frequencies across different activity classes. The formation of ACs is rationalized by relating structural similarity relationship and potency value distribution to each other, which are further analyzed, visualized and rationalized in an RMMP-based network representation.

- In *chapter 8*, activity class-dependent potency difference criteria are derived. Newly derived potency difference criteria are systematically compared with activity class-independent potency difference criteria (i.e., at least two orders of magnitude).

- In *chapter 9*, AC characteristics in analog series are investigated with activity class-dependent potency difference criteria. The overwhelming majority of ACs are single-site ACs. Multi-site ACs are also identified. Different potency effects of substituent combinations are observed, which are instructive for enlightening future compound design.

The final chapter (*chapter 10*) summarizes and discusses the main findings of this work.

Chapter 2

Simplified Activity Cliff Network Representations with High Interpretability and Immediate Access to SAR Information

Introduction

Activity cliffs (ACs) are formed by pairs of structurally similar compounds with large difference in potency. Although ACs are defined on the basis of compound pairs, more than 90% of ACs are actually formed in a coordinated manner rather than by isolated pairs of compounds. Coordinated ACs are formed by sets of structural analogs that participate in multiple cliffs. Such ACs can be organized and visualized in a network representation where nodes represent AC-forming compounds and edges pairwise AC relationships. In principle, coordinated ACs are more informative than isolated ACs. However, increasing numbers of densely connected AC compounds give rise to disjoint AC clusters of varying size and complexity, which complicates the immediate access to SAR information, Therefore efficient computational methods for deconvoluting AC clusters are required.

In this chapter, a new methodology for simplifying AC networks is introduced using three representative activity classes. The advantages of the novel approach are discussed.

Reprinted with permission from “Hu, H.; Bajorath, J. Simplified Activity Cliff Network Representations with High Interpretability and Immediate Access to SAR Information. *Journal of Computer-Aided Molecular Design* **2020**, *34*, 943-952”. *Copyright 2020 Springer Nature*

Journal of Computer-Aided Molecular Design (2020) 34:943–952
https://doi.org/10.1007/s10822-020-00319-9

Simplified activity cliff network representations with high interpretability and immediate access to SAR information

Huabin Hu[1] · Jürgen Bajorath[1]

Received: 3 April 2020 / Accepted: 29 May 2020 / Published online: 5 June 2020

Abstract

Activity cliffs (ACs) consist of structurally similar compounds with a large difference in potency against their target. Accordingly, ACs introduce discontinuity in structure-activity relationships (SARs) and are a prime source of SAR information. In compound data sets, the vast majority of ACs are formed by differently sized groups of structurally similar compounds with large potency variations. As a consequence, many of these compounds participate in multiple ACs. This coordinated formation of ACs increases their SAR information content compared to ACs considered as individual compound pairs, but complicates AC analysis. In network representations, coordinated ACs give rise to clusters of varying size and topology, which can be interactively and computationally analyzed. While AC networks are indispensable tools to study coordinated ACs, they become difficult to navigate and interpret in the presence of clusters of increasing size and complex topologies. Herein, we introduce reduced network representations that transform AC networks into an easily interpretable format from which SAR information in the form of R-group tables can be readily obtained. The simplified network variant greatly improves the interpretability of large and complex AC networks and substantially supports SAR exploration.

Key words Activity cliffs · Reduced activity cliff networks · SAR information · Matching molecular series · R-group tables

Introduction

Activity cliffs (ACs) are generally defined as pairs or groups of structurally similar or analogous compounds that share the same biological activity but have large differences in potency [1–3]. Accordingly, ACs encode small chemical changes having large effects on compound potency, which rationalizes their relevance for structure-activity relationship (SAR) analysis and chemical optimization [1–6]. For AC assessment, it must be decided when two compounds are sufficiently similar and their potency differences large enough to qualify as an AC. The evaluation of molecular similarity depends on chosen molecular representations and similarity measures [7]. For AC definition, different similarity and potency difference criteria are applicable and their choice characterizes different generations of ACs [8]. For systematic computational identification and analysis of ACs, consistent definitions must be applied [2, 3]. In addition, reliable AC assignments also depend on the use of high-quality activity measurements [6]. Much of our current knowledge about ACs and their distribution has resulted from systematic search calculations in large compound databases. Depending on the molecular representations that are used for structural similarity assessment and potency difference criteria that are applied, the frequency of ACs moderately varies. For example, ~ 20–30% of bioactive compounds participate in the formation of ACs and ~ 5–6% of pairs of structurally similar compounds form ACs if an at least 100-fold difference in potency is required [2, 3]. When alternative AC definitions are considered in parallel, on the order of 100,000 ACs are obtained on the basis of currently available bioactive compounds (unpublished data), which provide a rich source of SAR information.

One of the most important characteristics of ACs is that they rarely represent “isolated” compound pairs, i.e.,

Electronic supplementary material The online version of this article (doi:https://doi.org/10.1007/s10822-020-00319-9) contains supplementary material, which is available to authorized users.

✉ Jürgen Bajorath
bajorath@bit.uni-bonn.de

[1] Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Endenicher Allee 19c, 53115 Bonn, Germany

compounds having no other structural neighbors. Instead, ACs are typically formed by groups of structurally similar compounds with significant potency variations, giving rise to series of "coordinated" ACs in which many compounds are involved in multiple cliffs [9]. Regardless of the AC criteria that are applied, greater than 90% of all ACs found in compound activity classes are formed in a coordinated manner [9]. AC coordination can be explored in network representations, in which nodes represent compounds and edges pairwise ACs. In such networks, coordinated ACs give rise to the formation of AC clusters of varying size and topology [9]. AC clusters have higher SAR information content than ACs studied individually but, their interactive analysis is arduous when clusters increase in size and their topologies become rather complex [10]. Therefore, attempts have been made to computationally extract SAR information from AC clusters, for example, by organizing them in index maps on the basis of different intra-cluster structural relationships [10] or by isolating sequences of AC compounds from clusters that follow a potency gradient [11]. These approaches help to dissect clusters selected from AC networks and isolate AC subsets, providing at least partial access to SAR information.

While AC networks are essential for the rationalization and exploration of coordinated ACs, the interpretability of complex networks is limited. Difficulties in interpreting complex AC networks hinder SAR exploration on the basis of AC clusters. Therefore, we have developed a network variant that reduces complexity and provides immediate access to SAR information, as reported herein.

Materials and methods

Compound activity classes

Activity classes for AC network analysis were extracted from ChEMBL release 26 [12]. Compounds directly interacting with human targets (target relationship type: "D") at the highest assay confidence level (assay confidence score: 9) having equilibrium constants (K_i values) with exact "=" relationships as potency measurements were selected. If multiple measurements were available they were averaged, provided all potency values fell within the same order of magnitude; otherwise, the compound was disregarded. Table 1 summarizes the composition of three large activity classes used for AC network analysis.

Compound decomposition

Systematic single-cut fragmentation of exocyclic single bonds was carried out using an algorithm for the generation of matched molecular pairs (MMPs) [13]. An MMP is defined as a pair of compounds that are only distinguished by a chemical modification at a single site [13]. During each fragmentation step two fragments per compound were obtained including a core and a substituent. In the core, a hydrogen atom was added to the substitution site. Size restrictions were applied to confine cores and substituents to those typically observed in analog series [14]. First, the number of non-hydrogen (heavy) atoms in the core was required to be at least twice as large as in the substituent. Second, the substituent fragment was restricted to at most 13 heavy atoms. Third, the size difference between exchanged substituents in an MMP was set to at most eight heavy atoms.

Activity cliffs

For AC analysis, the MMP-cliff definition was used [14], which is tailored towards medicinal chemistry applications [6]. Accordingly, as AC criteria, two compounds from the same activity class are required to form a size-restricted MMP and have an at least 100-fold potency difference ($\Delta pK_i \geq 2.0$). By definition, MMP-cliffs contain a single substitution site.

Matching molecular series

As an extension of MMP concept, matching molecular series (MMSs) were systematically extracted from all AC compounds. An MMS consists of two or more analogs that share the same core (MMS-core) and are only distinguished by substituents at a single site [15]. All identified MMS-cores were subjected to a second round of MMP fragmentation, as described above, to identify structurally analogous cores.

Table 1 Activity classes

Target ID	Target name	No. CPDs	pK_i range	No. MMP-cliffs
259	Melanocortin receptor 4	1281	[3.65, 10.10]	426
244	Coagulation factor X	1641	[3.59, 11.40]	915
237	Kappa opioid receptor	1982	[4.09, 11.52]	987

For AC network analysis, three large activity classes were taken from ChEMBL. For each class, the ChEMBL target ID, target name, number of qualifying compounds (CPDs), their potency value (pK_i) range, and the number of MMP-cliffs are reported.

Two MMS-cores were structurally analogous if they formed a core-MMP and the corresponding MMSs were the classified as an MMS-pair (MMSP). Figure 1 shows an exemplary MMSP.

Networks

AC networks were generated in which nodes represent compounds and edges indicate the formation of pairwise MMP-cliffs [14]. Reduced AC networks were designed as detailed below. All network representations were drawn with Cytoscape [16].

Results and discussion

Network design principles

AC networks such as the one shown in Fig. 2 (top) are essential for visualizing and rationalizing the coordinated formation of ACs. Moreover, individual clusters emerging in AC networks provide a basis for the extraction of SAR information. With a total of 426 ACs (including only two isolated ACs) organized in 17 clusters, the AC network for melanocortin receptor 4 ligands has moderate size and complexity and is interpretable. However, extracting SAR information from the three largest clusters is already difficult, if not impossible by interactive analysis, requiring the application of computational approaches [10, 11]. We note that the use of the MMP concept as a substructure-based similarity criterion for AC formation supports interpretability of the network structure because MMP relationships are clearly defined and select structural analogs modified at a single site as AC compounds. Moreover, extension of the MMP concept through the MMS formalism makes it possible to trace MMSs in AC clusters as a basis for series-centric SAR analysis [11]. However, tracing single or multiple MMSs in AC clusters does not simplify the network structure [11].

To enable interpretation of AC networks of increasing size and complexity and facilitate direct extraction of SAR

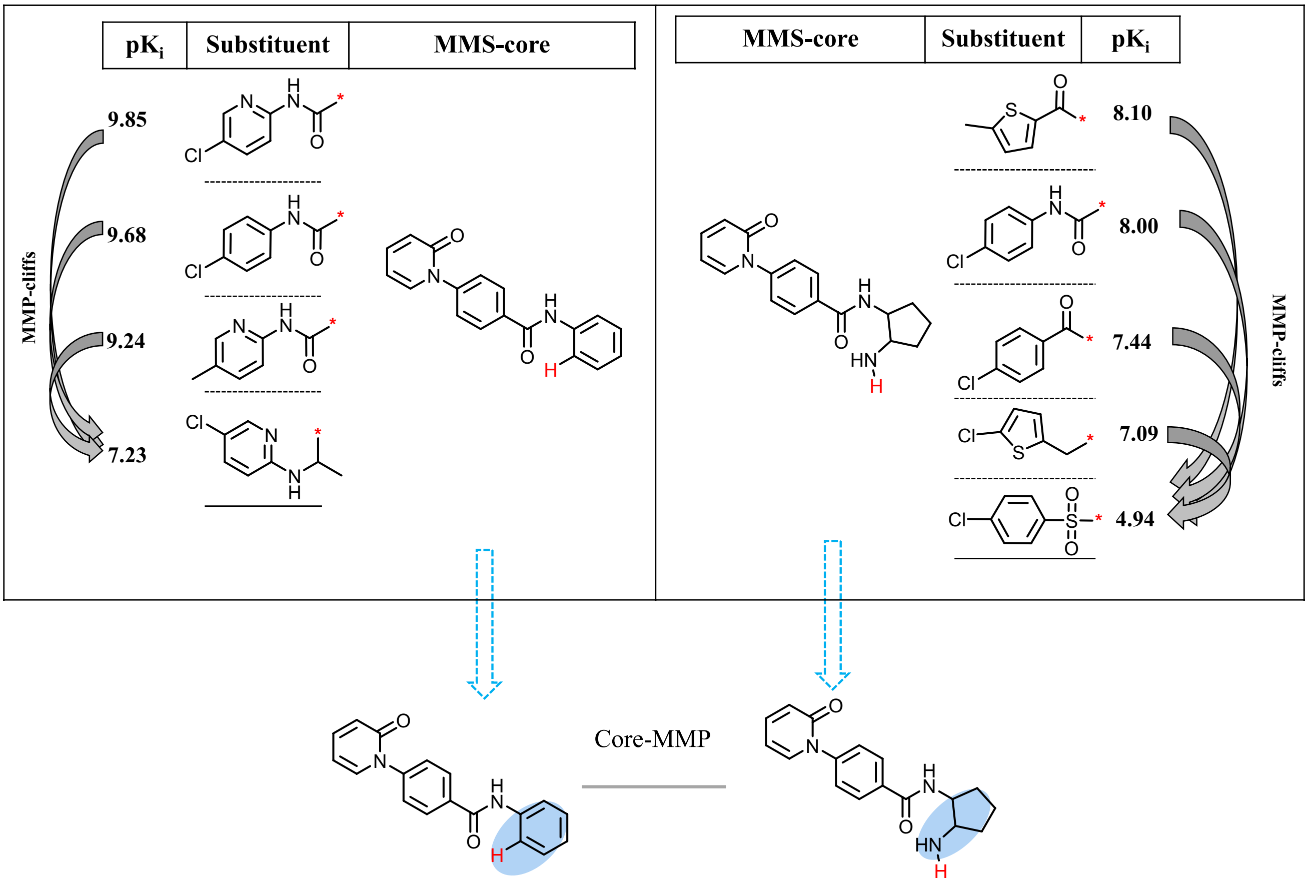


Fig. 1 Structural relationships. Shown are two MMSs of coagulation factor X inhibitors that contain multiple MMP-cliffs (indicated by curved arrows) and form an MMSP. MMSs are represented as R-group tables including compound potency (pKi) values. Hydrogen atoms added to the substitution sites in the two MMS-cores are colored red. The core-MMP resulting from the second round of fragmentation that establishes the relationships between these MMSs is shown at the bottom. Substituents distinguishing between MMS-cores are shown on a blue background.

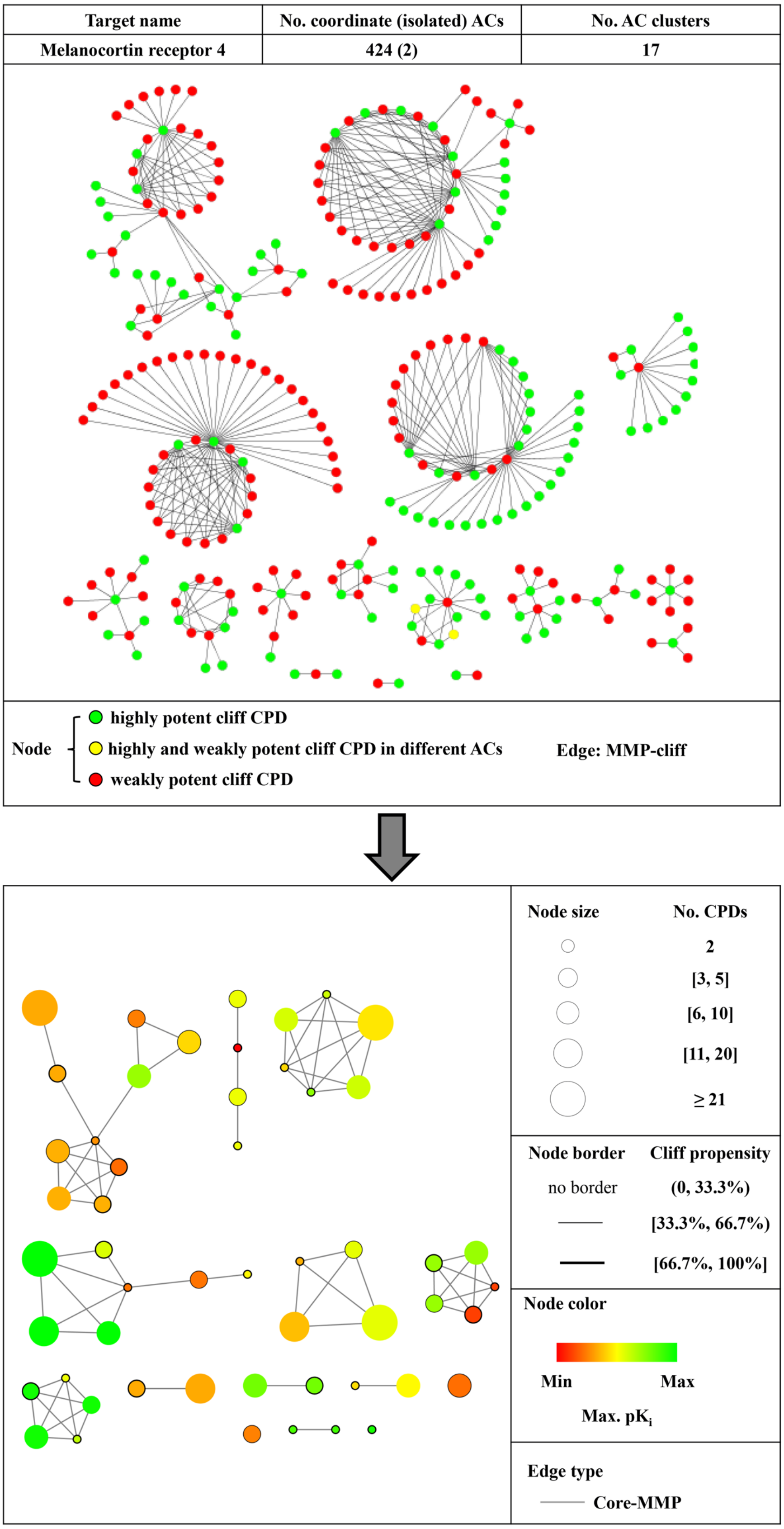


Fig. 2 Activity cliff network representations. At the top, the AC network formed by melanocortin receptor 4 ligands is shown that contains 424 coordinated and two isolated MMP-cliffs. Nodes represent AC compounds and edges the formation of pairwise MMP-cliffs. Nodes are color-coded to distinguish three types of AC compounds: green, highly potent AC compound; red, weakly potent AC compound; yellow, highly/weakly potent compound in different ACs. The network reveals the formation of AC clusters of varying size and topology. At the bottom, the reduced network is displayed. Design principles, as discussed in the text, are summarized on the right. In the reduced network, nodes represent MMSs and edges pairwise MMSP relationships

information, we have developed an approach for the reduction of AC networks that employs the MMS formalism in different ways. Design principles for the simplified network are summarized in Fig. 2 (bottom). A central idea underlying the network reduction approach is transforming the entire cluster structure of the AC network into an array of MMSs and MMSPs. Thereby, all ACs are represented on the basis of MMSs and structurally related MMS-cores are identified. In the corresponding reduced network, each node represents an MMS comprising two or more analogs. The inclusion of compound pairs accounts for isolated ACs. Edges between nodes indicate MMSP relationships (in algorithmic terms, the formation of a core-MMP). Nodes are scaled in size according to the number of compounds per MMS and can be color-coded according to different potency characteristics (or other properties) such as the largest potency of MMS members. This color scheme accounts for the distribution of highly potent AC compounds across MMSs. AC information is also conveyed through node borders, the thickness of which reflects the AC propensity within MMSs. Propensity represents the percentage of all possible analog pairs that form an AC in a given MMS. By design, individual MMS clusters in the reduced network may combine multiple original AC clusters, but have simpler topologies and limited complexity. However, all AC information is retained and MMSs or MMSPs with high AC propensity can be readily identified and selected for further analysis.

Reducing complex activity cliff networks

The utility of reduced networks becomes immediately evident when AC networks of increasing size and complexity are considered such as the example in Fig. 3a. The network at the top consists of 915 ACs (including only 15 isolated ACs) and contains several densely connected spherical clusters. The two largest AC clusters are essentially impossible to analyze interactively. By contrast, the reduced network at the bottom is immediately interpretable. It consists of 91 MSSs including 71 that form a total of 87 MMSPs. In addition, there are 20 single MSSs. In the reduced network, the largest AC cluster (with 363 ACs) from the original network is mostly (96%) represented by the MMS cluster encircled using a blue dashed line. It can be seen that this cluster combines nine MMSs of greatly varying size that contain highly potent cliff compounds. Seven of the nine MMSs are densely connected including the two largest and the smallest ones. The remaining two MMSs only form one or two pairs including a medium size MMS with multiple ACs. In contrast to the original AC network, the reduced network can be easily navigated including the largest clusters. Another example is shown in Fig. 4a. Here, the AC network of kappa opioid receptor ligands (top) comprises 987 ACs that are organized in 54 clusters, the largest of which dominates the network view. In the reduced network (bottom), this very large and densely connected cluster (with 493 ACs) is exclusively represented by the encircled MMS cluster at the upper left (containing MMSP 1/2). Other clusters in the reduced network have simple topologies and are straightforward to analyze.

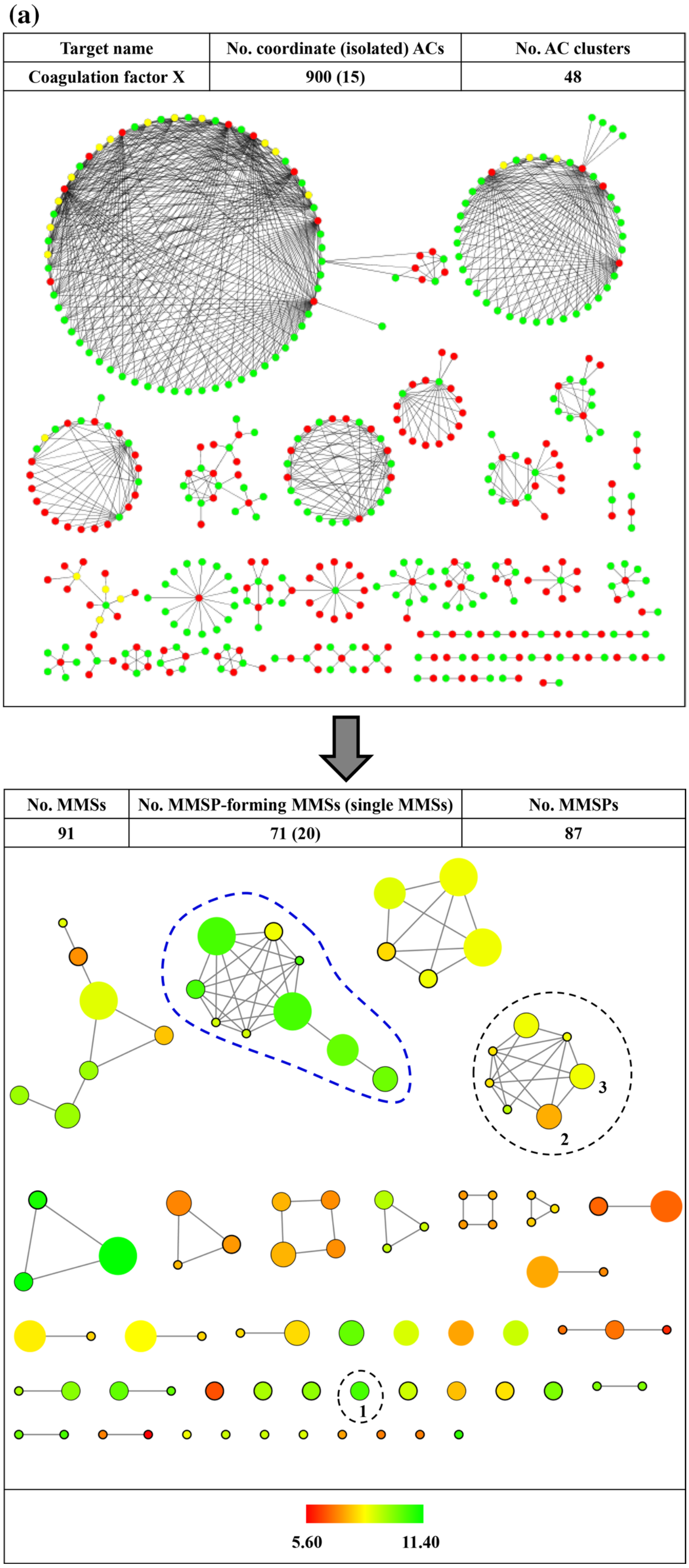
Extracting SAR information from reduced networks

A key feature of reduced networks is that individual MMSs or MMSPs of interest can be easily selected and represented in standard R-group tables. These tables are most popular in medicinal chemistry for the representation of analog series and provide immediate access to SAR information including ACs formed within the MMSs. Examples are shown in Figs. 3b and 4b. Compared to original AC networks, extraction of SAR information from reduced networks is greatly simplified. Notably, generating R-group tables from MMSPs, as shown in Figs. 3b and 4b, further supports SAR analysis compared to single MMSs. This is the case because MMS-cores of MMSPs are structurally analogous by design. Since these cores are algorithmically generated for large-scale AC analysis, they should always be compared from a chemical perspective when individual MMSs are considered. In some instances, algorithmically defined cores might be chemically sufficiently similar such that the R-group tables of the MMSP can be jointly analyzed or even combined. For example, this would be the case for the MMSP in Fig. 3b. In other instances, cores might be chemically distinct -although they are structurally analogous- likely giving rise to different SAR characteristics exhibited by related MMSs. Examples are provided in Fig. 4b. Since these MMSs from reduced networks contain ACs, they likely reveal SAR determinants for related yet distinct series. The reduced networks provide many opportunities for comparing SARs encoded by MMSPs on the basis of their R-group tables, which benefits SAR exploration from a medicinal chemistry perspective.

Conclusions

The vast majority of ACs are formed in a coordinated manner. For their analysis, network representations play a central role. In AC networks, coordinated ACs centred on different analog series emerge as disjoint clusters of different composition and varying topology. These AC clusters become a primary focal point for SAR exploration. However, with increasing size and complexity, AC networks become difficult to navigate and clusters hard to analyze interactively. Accordingly, there is a need for making coordinated ACs and the information they provide available in a format that is readily interpretable. We have reasoned that network reduction might be suitable for this purpose, provided that

Fig. 3 Activity cliff networks for coagulation factor X inhibitors. In **a**, the original AC network (top) and the reduced network (bottom) are displayed according to Fig. 2. Numbers at an encircled node and cluster mark an exemplary isolated MMS (1) and an MMSP (2/3), respectively. In **b**, R-group tables representing the isolated MMS (top) and MMSP (bottom) are shown.



Target name	No. coordinate (isolated) ACs	No. AC clusters
Coagulation factor X	900 (15)	48

No. MMSs	No. MMSP-forming MMSs (single MMSs)	No. MMSPs
91	71 (20)	87

Fig. 3 (continued)

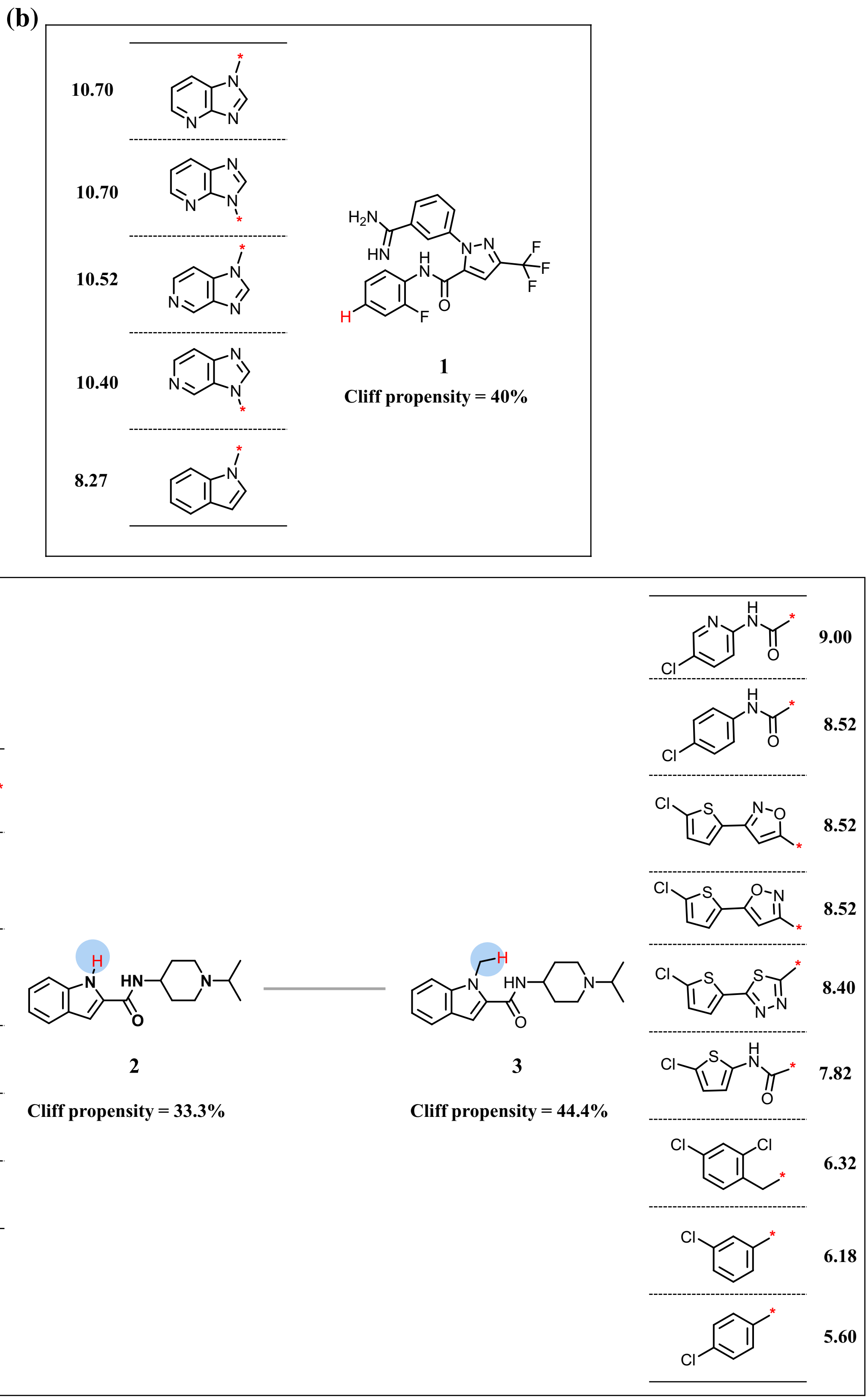
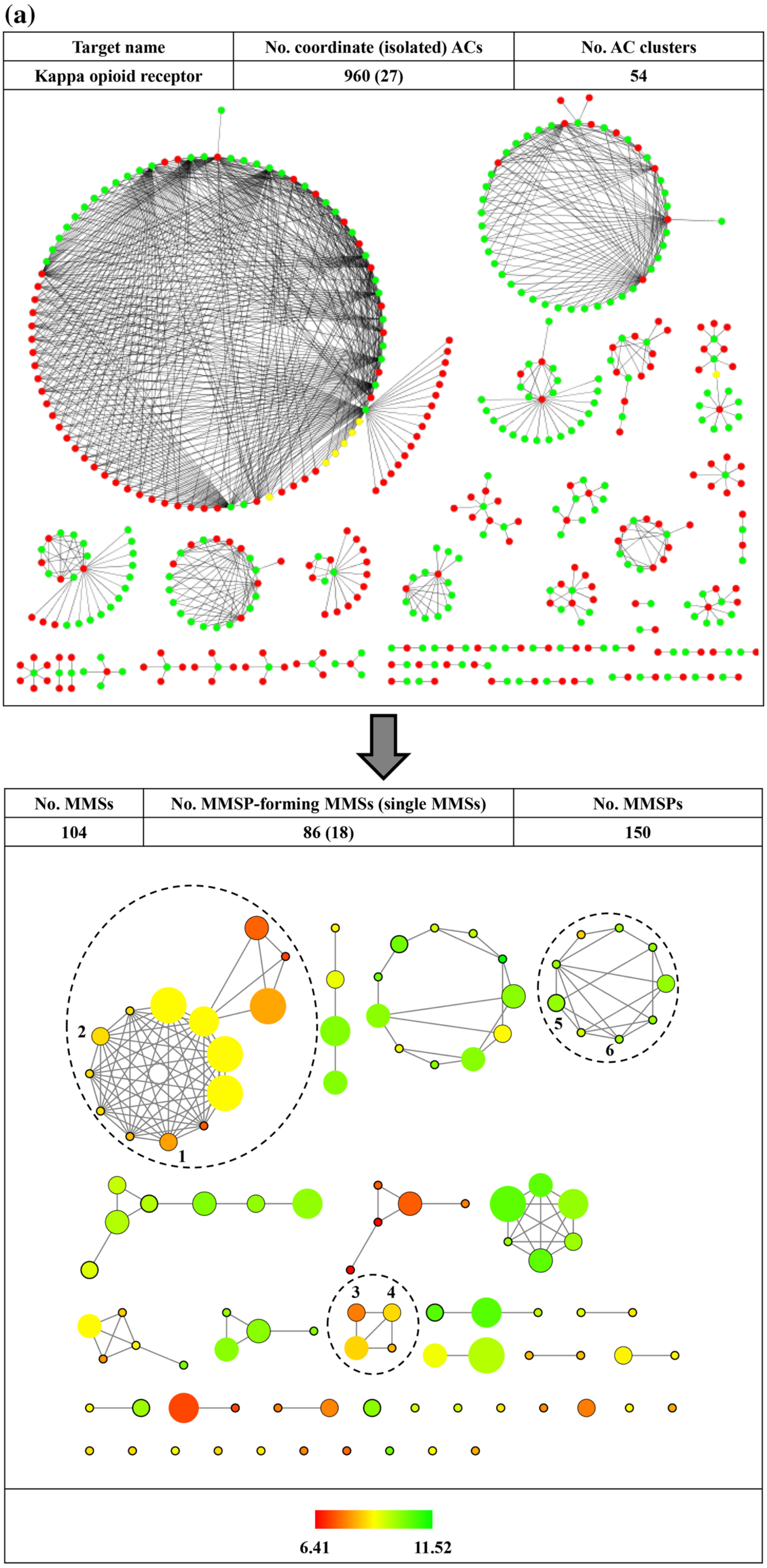


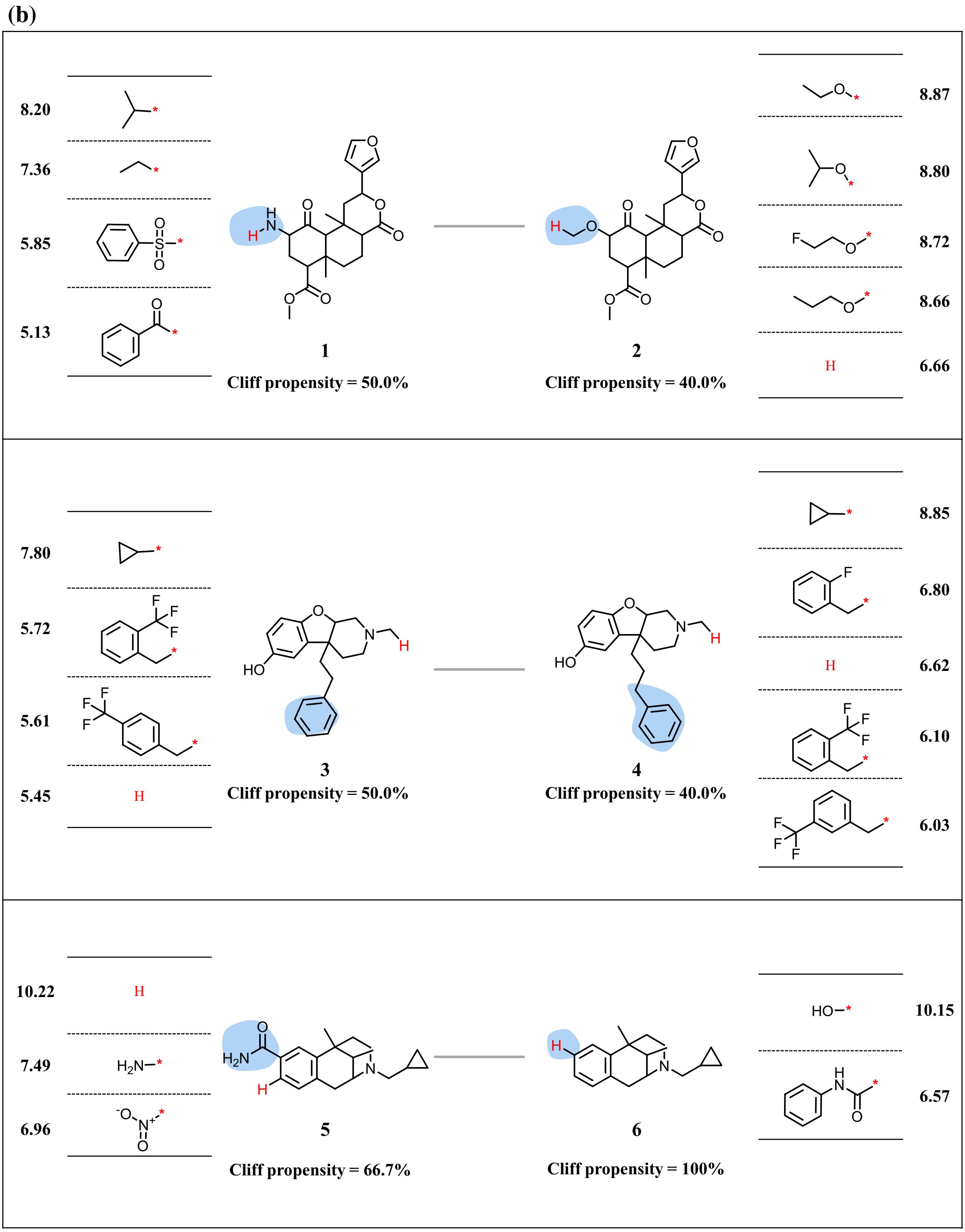
Fig. 4 Activity cliff networks for kappa opioid receptor. In **a** the original AC network (top) and the reduced network (bottom) are displayed according to Fig. 2. Numbers at encircled clusters mark three exemplary MMSPs (1/2, 3/4, and 5/6). In **b** R-group tables representing the three MMSPs are shown



Target name	No. coordinate (isolated) ACs	No. AC clusters
Kappa opioid receptor	960 (27)	54

No. MMSs	No. MMSP-forming MMSs (single MMSs)	No. MMSPs
104	86 (18)	150

Fig. 4 (continued)



AC information could be fully retained. Therefore, in this work, we have introduced an approach for the generation of simplified AC networks that is conceptually based upon the MMS formalism and the assessment of structural relationships between MMSs. In reduced networks, resulting MMSPs and individual MMSs resolve the original AC cluster structure and replace it with a higher-level structural organization scheme, which results in simplified network views and ensures interpretability. This represent a key aspect of the design strategy. As shown herein, original and reduced networks can be analyzed side-by-side, providing complementary views. Moreover, from reduced networks, MMSs and MMSPs can be easily selected and represented as R-group tables that reveal ACs and SAR information. This is another key feature of the approach. Presenting analog series from simplified networks in the form of R-group tables enables SAR analysis from a medicinal chemistry perspective, without requiring further computational input, and hence supports practical applications. In our proof-of-concept study, representative activity classes and AC populations have been investigated to demonstrate the utility of the approach. Reduced networks have been generated for many more activity classes, consistently enabling interpretation of AC clusters and SAR analysis on the basis of R-group tables. We also note that reduced network representations will not replace original AC networks, but are designed to aid in their analysis through the generation of complementary simplified views. AC networks remain important tools for globally visualizing the coordinated formation of ACs and comparing AC populations originating from different compound data sets. However, reduced networks will be essential for detailed analysis of large AC clusters with complex topologies.

Acknowledgements Open Access funding provided by Projekt DEAL.

References

1. Maggiora GM (2006) On outliers and activity cliffs - why QSAR often disappoints. J Chem Inf Model 46:1535–1535
2. Stumpfe D, Bajorath J (2012) Exploring activity cliffs in medicinal chemistry. J Med Chem 55:2932–2942
3. Stumpfe D, Hu Y, Dimova D, Bajorath J (2014) Recent progress in understanding activity cliffs and their utility in medicinal chemistry. J Med Chem 57:18–28
4. Medina-Franco JL (2013) Activity cliffs: facts or artifacts? Chem Biol Drug Des 81:553–556
5. Cruz-Monteagudo M, Medina-Franco JL, Pérez-Castillo Y, Nicolotti O, Cordeiro MNDS, Borges F (2014) Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? Drug Discov Today 19:1069–1080
6. Stumpfe D, Hu H, Bajorath J (2019) Evolving concept of activity cliffs. ACS Omega 4:14360–14368
7. Maggiora G, Vogt M, Stumpfe D, Bajorath J (2014) Molecular similarity in medicinal chemistry. J Med Chem 57:3186–3204
8. Stumpfe D, Hu H, Bajorath J (2020) Computational method for the identification of third generation activity cliffs. MethodsX 7:e100793
9. Stumpfe D, Dimova D, Bajorath J (2014) Composition and topology of activity cliff clusters formed by bioactive compounds. J Chem Inf Model 54:451–461
10. Dimova D, Stumpfe D, Bajorath J (2014) Method for the evaluation of structure-activity relationship information associated with coordinated activity cliffs. J Med Chem 57:6553–6563
11. Dimova D, Bajorath J (2014) Extraction of structure-activity relationship information from activity cliff clusters via matching molecular series. Eur J Med Chem 87:454–460
12. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP (2012) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40:D1100–D1107
13. Hussain J, Rea C (2010) Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. J Chem Inf Model 50:339–348
14. Hu X, Hu Y, Vogt M, Stumpfe D, Bajorath J (2012) MMP-cliffs: systematic identification of activity cliffs on the basis of matched molecular pairs. J Chem Inf Model 52:1138–1145
15. Wawer M, Bajorath J (2011) Local structural changes, global data views: graphical substructure-activity relationship trailing. J Med Chem 54:2944–2951
16. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2020) Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27:431–432

Summary

Coordinated ACs forming clusters of varying size contain much SAR information. A variety of approaches have been developed to systematically explore SAR information in AC clusters, for instance, by defining numerical indices indicating intra-cluster structural relationships and AC diversity, or by identifying compounds shared between two MMSs to analyze SARs in a pairwise manner. Herein we introduce another methodology aiming to simplify the analysis of complex AC clusters. In the original MMP-cliff network, one disjoint AC cluster contains multiple compounds with one or several different MMP cores which can be organized into distinct MMSs according to the shared MMP cores. Then a second-round core fragmentation is conducted to build edge relationships between individual MMSs. Therefore, in the reduced network, the nodes represent individual MMSs instead of individual compounds as in AC networks and the edges MMS pairs for which the cores form MMP relationships. Additional node properties such as mean potency values of the AC compounds for each MMS, number of AC compounds, or cliff propensity of each MMS can be added and easily visualized in the reduced network.

After AC network simplification, the SAR information of large AC clusters is intuitively interpretable and easily accessible. In the next chapter, we extend the MMP-cliff data structure with the inclusion of structural isomers, leading to the introduction of isomer/MMP-cliffs.

Chapter 3

Introducing a New Category of Activity Cliffs Combining Different Compound Similarity Criteria

Introduction

Given the popularity of MMP-based compound similarity evaluations, MMP-cliffs have been intensively studied. By definition, two compounds forming an MMP relationship are distinguished at a single substitution site. Accordingly, the formation of an MMP-cliff indicates the replacement of one substituent (R-group) with another one at the same core position resulting in a large difference in potency. A natural question to ask is whether the replaced R-group has to be confined to a designated core position to elicit a strong biological effect. R-groups attached at different core positions are known as structural isomers in organic chemistry.

In this chapter, we systematically extracted MMP-cliffs and identified structural isomers for MMP-cliff compounds carrying the same substituent at varying core positions. Potency values of structural isomers and MMP-cliff compounds are compared giving rise to the new AC category: isomer/MMP-cliffs.

RSC
Medical Chemistry

ROYAL SOCIETY OF CHEMISTRY

RESEARCH ARTICLE

View Article Online
View Journal

Check for updates

Cite this: DOI: 10.1039/c9md00463g

Introducing a new category of activity cliffs combining different compound similarity criteria

Huabin Hu and Jürgen Bajorath*

Received 27th September 2019,
Accepted 21st December 2019

DOI: 10.1039/c9md00463g

rsc.li/medchem

Activity cliffs (ACs) are pairs of structurally similar or analogous active compounds with large differences in potency against the same target. For identifying and analyzing ACs, similarity and potency difference criteria must be determined and consistently applied. This can be done in various ways, leading to different types of ACs. In this work, we introduce a new category of ACs by combining different similarity criteria, including the formation of matched molecular pairs and structural isomer relationships. A systematic computational search identified such ACs in compounds with activity against a variety of targets. In addition to other ACs exclusively formed by structural isomers, the newly introduced category of ACs is rich in structure–activity relationship (SAR) information, straightforward to interpret from a chemical perspective, and further extends the current spectrum of ACs.

Introduction

Structurally similar active compounds with large potency differences form activity cliffs (ACs).[1,2] They can be detected in analog series during chemical optimization or extracted from compound data sets. ACs reveal small chemical modifications that significantly impact biological activity and are thus of high interest in structure–activity relationship (SAR) analysis.[2,3] In the practice of medicinal chemistry, ACs might be subjectively assessed on a case-by-case basis when encountered during compound optimization efforts. However, for systematic identification and organization as well as consistent representation and evaluation of ACs, similarity and potency difference criteria must be clearly defined and consistently applied.[2,3] We note that similarity is generally considered as a subjective criterion but in chemistry and other scientific fields, different metrics and measures have been introduced to quantify similarity in reproducible ways.[4] In medicinal chemistry, this provides the foundation for establishing compound similarity relationships beyond subjective assessment and chemical intuition and enabling systematic SAR exploration.[4] For large-scale identification and analysis of ACs, computational methods play an important role.[2] The choice and combination of alternative similarity and potency difference criteria give rise to different categories of ACs having different characteristics.

The question if two compounds are sufficiently similar to form an AC can be addressed in different ways, for example, by calculating Tanimoto similarity on the basis of graph-based molecular representations or by applying substructure-based similarity concepts.[2–4] Substructure-based measures include, for example, the conservation of compound scaffolds,[4,5] formation of matched molecular pairs (MMPs),[6,7] or presence of analog relationships (*i.e.*, two compounds belong to the same analog series).[8,9] If numerical similarity metrics are applied, a similarity threshold for AC formation must be set, which is not only representation-dependent, but also subjective in nature.[3]

For substructure-based similarity assessment, MMPs have become increasingly popular. They are defined as pairs of compounds that only differ by a confined chemical change at a single site, which is termed a chemical transformation.[6] MMPs can be extracted from large compound collections in computationally efficient ways,[6] which supports large-scale analysis. Hence, the MMP concept can also be applied to computationally identify structural analogs[7] and series of analogs.[8] In addition to applying molecular graph-based similarity measures, ACs have also been determined on the basis of X-ray structures of ligand-target complexes.[10] This requires the calculation of three-dimensional (3D) similarity of experimentally observed compound binding modes, yielding so-called 3D-cliffs.[10]

The question when potency differences between analogs become sufficiently large to qualify a compound pair as an AC can also be addressed in different ways. For example, a constant potency difference threshold can be applied that reflects a statistically significant difference in potency across many compound data sets (activity classes, also termed

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Endenicher Allee 19c, D-53115 Bonn, Germany. E-mail: bajorath@bit.uni-bonn.de; Fax: +49 228 7369 100; Tel: +49 228 7369 100

target sets).[2,3] Alternatively, target set-dependent potency difference thresholds can be determined, which take set-specific potency value distributions into account.[11] As a constant potency difference threshold, a potency difference of at least two orders of magnitude between similar candidate compounds has often been applied.[3,7] For comparison, potency difference thresholds of target set-dependent ACs frequently range from 1.5 to 2.5 orders of magnitude[11] and are thus comparable.

The application of different similarity measures and potency difference thresholds characterize subsequent generations of ACs, beginning with ACs that were defined on the basis of numerical similarity measures and constant potency difference thresholds[1,2] and leading to ACs based upon substructure-based similarity criteria and target set-dependent potency difference thresholds.[9,11]

For large scale-analysis of ACs across many different compound classes, MMP-cliffs[6] have been particularly useful,[12,13] given their computationally efficient generation and chemically intuitive nature. For MMP-cliffs, the similarity criterion that must be met by candidate compounds is the formation of a transformation size-restricted MMP and a constant potency difference of at least two orders of magnitude is required.[7]

Herein, we introduce a new category of ACs by assessing similarity in a previously unconsidered manner. For the first time, different similarity criteria are applied in combination to define ACs, leading to the identification of new ACs with high SAR information content for a variety of pharmaceutical targets.

Materials and methods

Compounds and activity data

Bioactive compounds were extracted from ChEMBL version 24.1.[14] For our analysis, the following selection criteria were applied. Only compounds with direct interactions (target relationship type "D") with human target proteins at the highest assay confidence level (ChEMBL confidence score 9) and available numerically specified equilibrium constants (K_i values) were selected. Approximate measurements such as those indicated by "<", ">" or "~" were not considered. On the basis of these criteria, a total of 73 965 unique compounds with activity against 915 targets were obtained and divided into 915 target sets.

Systematic compound fragmentation

Following the MMP fragmentation scheme,[6] exocyclic bonds in test compounds were subjected to systematic single-cut fragmentation (*i.e.*, a single bond was cleaved per iteration), which produced two substructures (core and substituent). The following size restrictions were applied:[7] the size of the core (number of non-hydrogen atoms) was required to be at least twice the size of the substituent and the size of the substituent was limited to at most 13 non-hydrogen atoms. The fragmentation protocol was applied to systematically generate MMPs and identify structural isomers (see below).

Generation of matched molecular pairs

An MMP is defined as a pair of compounds that only differ by a chemical change at a single site. For MMP generation, the size restrictions specified above were complemented by applying an additional rule, *i.e.*, the size difference between exchanged fragments (representing a chemical transformation) was limited to at most eight non-hydrogen atoms. The application of these rules yielded transformation size-restricted MMPs.[7]

Identification of structural isomers

Structural isomers are compounds that have the same chemical composition formula but are topologically distinct. Herein, structural isomers were identified that only differed in the core position of the substituent fragment, corresponding to sets of analogs in which the same substituent fragment occurred at different positions. To systematically identify and classify such structural isomers in target sets, generalized cores were constructed with the aid of the OpenEye Chemistry toolkit,[15] in which each attachment site of a substituent fragment was substituted with a hydrogen atom. All structural isomers originating from a target set that were represented by the same generalized core and fragment were then combined into an isomer set.

Activity cliff criteria

Three types of ACs were investigated herein, consistently requiring an at least 100-fold difference in potency between cliff compounds. First, standard MMP-cliffs were extracted from target sets. In addition, "isomers cliffs" were defined to be formed by two structural isomers from the same set, also having an at least 100-fold difference in potency. A separate search for isomer cliffs was carried out. We note that isomer cliffs, as defined herein, are related to "topology cliffs" that were reported previously applying a scaffold-based similarity criterion.[4] Furthermore, "isomer/MMP-cliffs" were introduced. As described in more detail below, in isomer/MMP-cliffs, one MMP-cliff compound was replaced by a structural isomer. Hence, searching for isomer/MMP-cliffs required combining MMP- and structural isomer-based similarity assessment. Therefore, as a pre-requisite of identifying isomer/MMP-cliffs, MMP-cliff compounds were determined that also belonged to isomer sets. The corresponding ACs were termed "MMP-cliffs with isomer extension". If multiple MMP-cliffs were found to be associated with the same isomer set, only the MMP-cliff with the largest potency difference was retained for further exploration of isomer/MMP-cliffs, thus avoiding potential AC redundancy.

Results and discussion

Systematic exploration of structural relationships

Fig. 1 illustrates different structural relationships investigated in this work. Bioactive compounds with high-confidence activity data were systematically searched for transformation size-restricted MMPs. In parallel, a search was carried out for sets of structural isomers. Then, MMP-cliff compounds were identified that also participated in isomer sets, thus combining MMPs and isomer sets for AC analysis.

Extending the current spectrum of activity cliffs

Three types of ACs investigated herein are depicted in Fig. 2. As a standard, MMP-cliffs were systematically identified. In addition, isomer sets were independently identified and searched for pairs of isomers with an at least 100-fold difference in potency, yielding isomer cliffs. In these ACs, compounds were distinguished by the position of a given substituent (resulting from molecular fragmentation). So-called "chirality cliffs"[5] or "chiral cliffs"[16] in which compounds with large potency differences are only distinguished by the configuration at a single stereo center have been described previously.[5,16] By contrast, isomer cliffs as defined herein have not been introduced before (but -as stated above- are related to scaffold-based topology cliffs). Moreover, isomer/MMP-cliffs also shown in Fig. 2 represent a novel category of ACs. We reasoned that adding structural isomers to MMP-cliffs would further extend their SAR information content. By definition, structural analogs forming MMP-cliff are distinguished by a substitution at one and only one site. However, replacing an MMP-cliff compound by a structural isomer adds another substitution site. Hence, compounds forming an isomer/MMP-cliff are distinguished by different substituents (R-groups) at two sites, which can be accounted for following MMP terminology as H ↔ R transformations. Combining different similarity criteria is a characteristic feature of isomer/MMP-cliffs setting them apart from other AC categories.

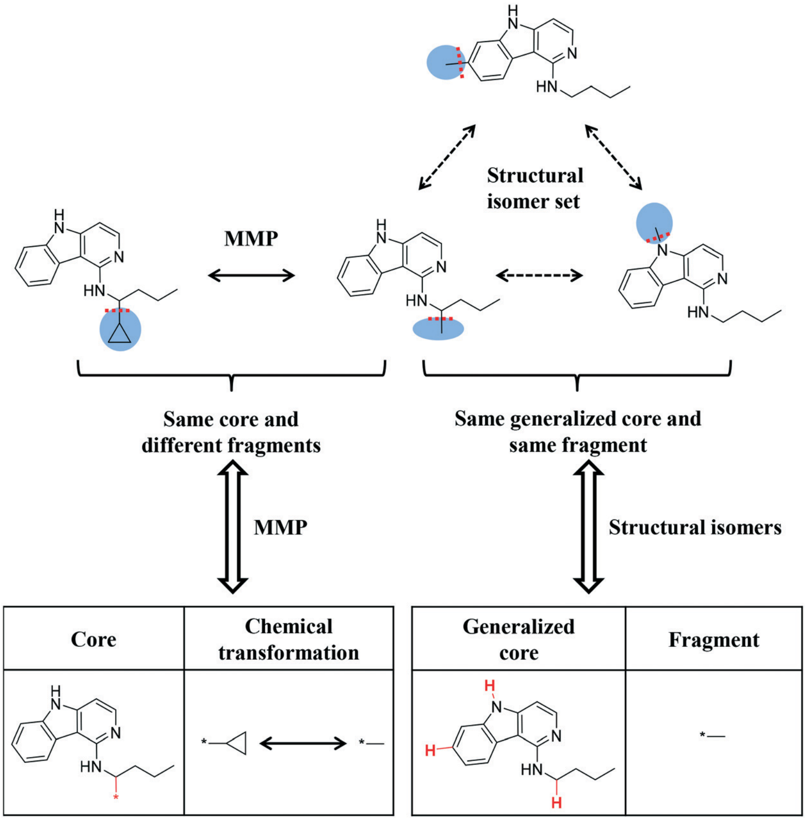


Fig. 1 Structural relationships. The schematic representation illustrates structural relationships that were systematically identified. For this purpose, a small compound (CPD) set with four analogs is used. For this exemplary set, MMP and structural isomer relationships are shown. Initially, compounds from all qualifying target sets were subjected to fragmentation of exocyclic single bonds to detect MMPs. Cleaved bonds are indicated by dashed red lines and the resulting fragments are shown on a blue background. For each MMP, the chemical transformation was recorded. The MMP fragmentation scheme was also adapted to identify structural isomers (that share the same composition formula, but are topologically distinct). Therefore, a search was carried out for structurally distinct (unique) compounds that yielded the same fragment and core of the same composition. Such compounds were represented by the same fragment and generalized core, in which fragmentation sites were hydrogen substituted (shown in red), and combined into an isomer set. Finally, MMP compounds were identified that also belonged to isomer sets, thereby combining different structural relationships.

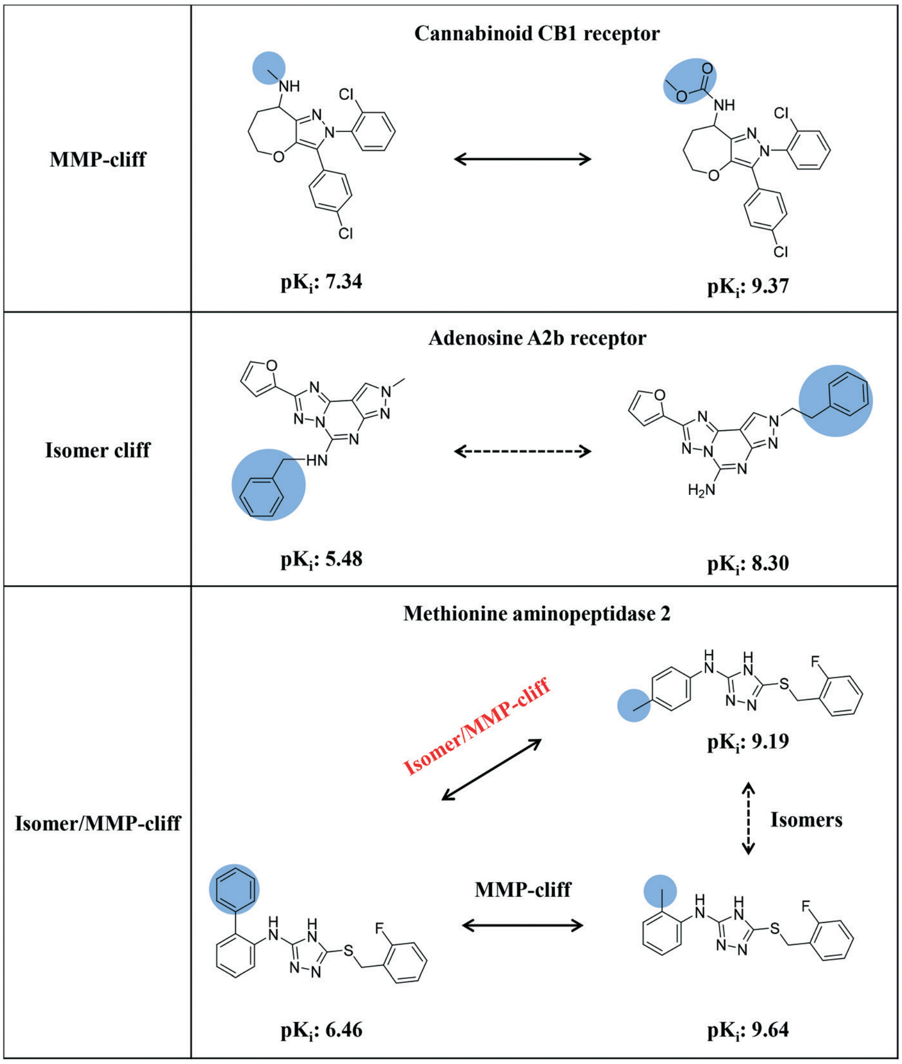


Fig. 2 Activity cliff categories. Shown are exemplary ACs belonging to different categories including (from the top to the bottom) an MMP-cliff, isomer cliff, and isomer/MMP-cliff. In each case, the target of the AC compounds is given.

Searching for activity cliffs

A systematic search for the three types of ACs was carried out in ChEMBL, as summarized in Fig. 3. From nearly 74 000 compounds with qualifying activity data for 915 targets, more than 600 000 MMPs were extracted that yielded 26 966 MMP-cliffs originating from 351 target sets. These MMP-cliffs involved 14 008 unique compounds. In addition, 10 571 different isomer sets (with the median and maximum size of two and eight isomers, respectively) were identified comprising 13 867 unique compounds from 412 target sets. These isomer sets contained 16 314 isomer pairs that yielded 493 isomer cliffs for 124 different targets. Fig. 4a shows that in only 425 (4.0%) of all isomer sets, the potency difference threshold for AC formation was met. By contrast, in 5706 isomer sets, maximal pairwise potency differences were close to zero and in more than 8000 sets, they fell within one order of magnitude. Hence, structural isomers mostly had similar potency against their targets. Although the absolute number of isomer cliffs was much smaller than of MMP-cliffs, the percentage of isomer cliffs among isomer pairs (3.0%) was comparable to the proportion of MMP-cliffs among MMPs (4.4%). Fig. 4b shows two isomer sets with three isomers each in which isomer cliffs were formed.

Next, we searched for MMP-cliffs with isomer extension, *i.e.*, MMP-cliff compounds that also belonged to isomer sets. For 1182 MMP-cliffs (4.4%) originating from 147 target sets, structural isomers were identified, as reported in Fig. 3. From these MMP-cliffs with isomer extension, a total of 597 isomer/MMP-cliffs were extracted, which consisted of 636 unique compounds with activity against 80 different targets. Thus, 39.8% of MMP-cliffs with isomer extension represented isomer/MMP cliffs and provided informative AC constellations for further analysis, as discussed below. First, we take a closer look at MMP-

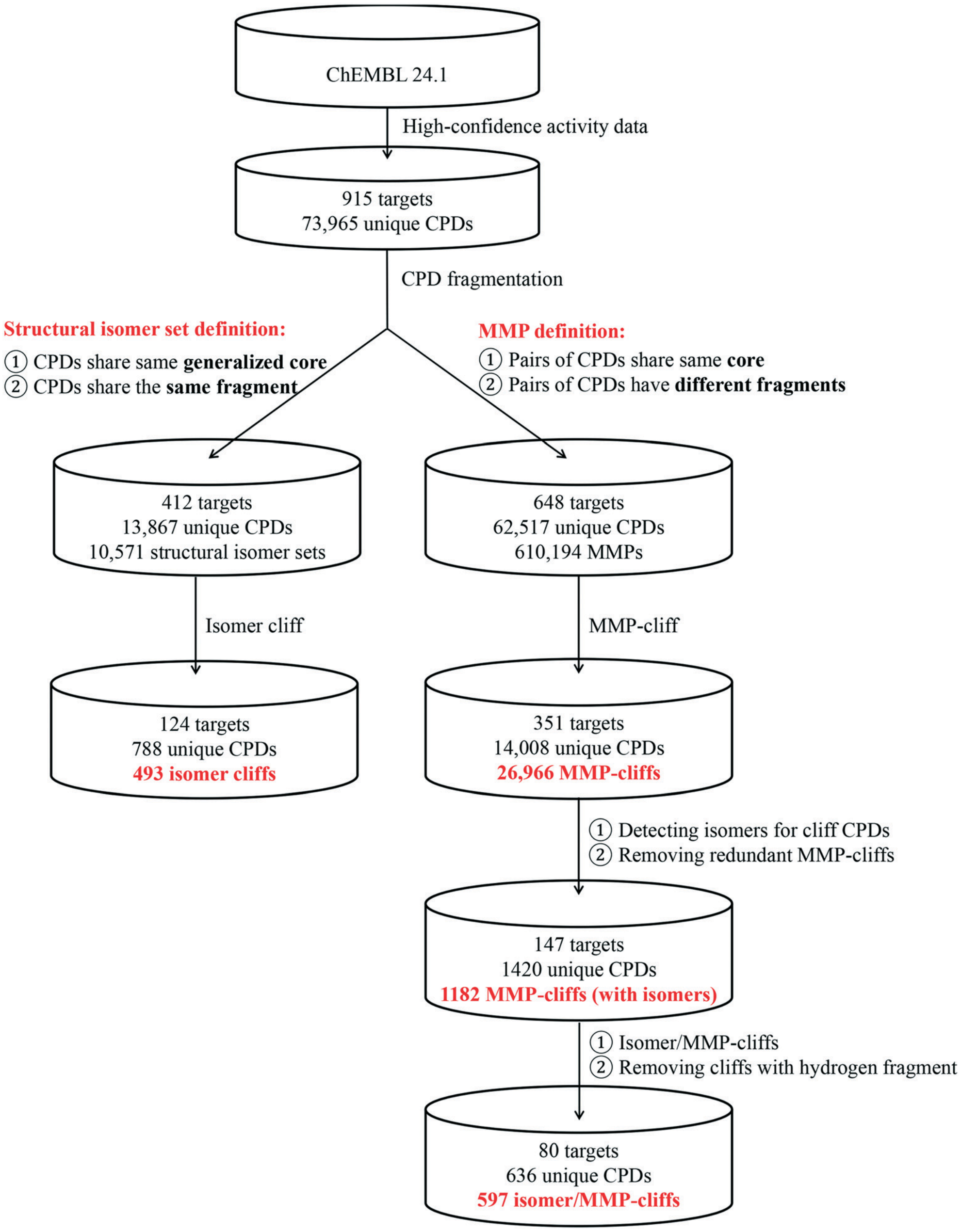

ChEMBL 24.1
High-confidence activity data
915 targets
73,965 unique CPDs
CPD fragmentation
Structural isomer set definition:
① CPDs share same generalized core
② CPDs share the same fragment
MMP definition:
① Pairs of CPDs share same core
② Pairs of CPDs have different fragments
412 targets
13,867 unique CPDs
10,571 structural isomer sets
648 targets
62,517 unique CPDs
610,194 MMPs
Isomer cliff
MMP-cliff
124 targets
788 unique CPDs
493 isomer cliffs
351 targets
14,008 unique CPDs
26,966 MMP-cliffs
① Detecting isomers for cliff CPDs
② Removing redundant MMP-cliffs
147 targets
1420 unique CPDs
1182 MMP-cliffs (with isomers)
① Isomer/MMP-cliffs
② Removing cliffs with hydrogen fragment
80 targets
636 unique CPDs
597 isomer/MMP-cliffs

Fig. 3 Identification of different activity cliffs. The workflow chart summarizes the identification of MMP-cliffs, isomer cliffs, MMP-cliffs with isomer extension, and isomer/MMP-cliffs across different target sets.

cliffs with isomer extension and the chemical transformations they contained.

Chemical transformations in extended MMP-cliffs

We systematically analyzed chemical transformations associated with MMP-cliffs. For the 1182 MMP-cliffs with isomer extension, 676 unique chemical transformations were detected. Interestingly, small transformations involving hydrogen atom replacements were among the most frequently observed chemical changes in MMP-cliffs with isomer extension, as shown in Fig. 5a. The replacement of a hydrogen atom by a methyl group occurred most frequently (and with similar frequency as in all MMP-cliffs), followed by hydrogen replacements with a methoxy group and chlorine atom, respectively. For extended MMP-cliffs with hydrogen atom replacements, broad distributions of maximum potency differences between MMP-cliff compounds and structural isomers were observed, as reported in Fig. 5b, often with median values around two orders of magnitude.

Fig. 6a shows exemplary MMP-cliffs with an H ⟷ CH_3 transformation for which structural isomers of weakly potent cliff compounds were available. These extended MMP-cliffs

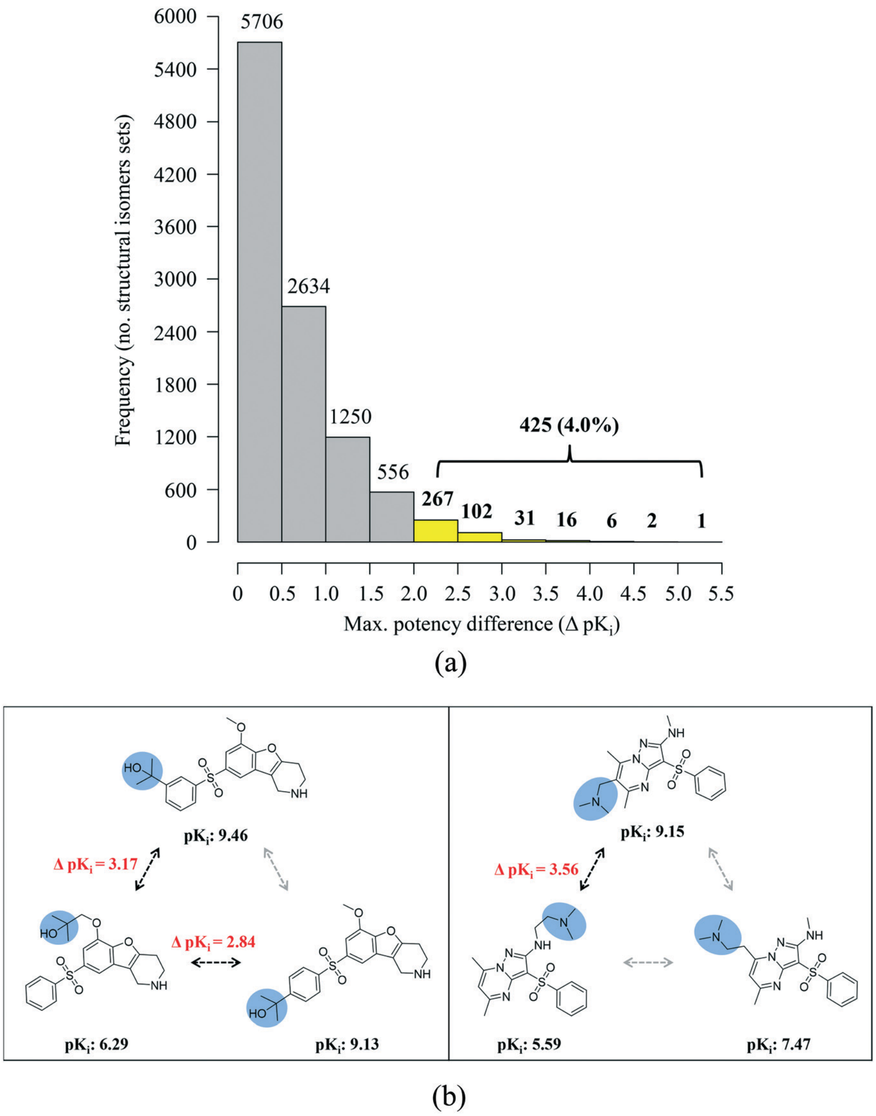


Fig. 4 Potency differences in isomer sets and isomer cliffs. (a) The distribution of maximum pairwise potency differences (ΔpK_i) in isomer sets is reported. In only 4% of all isomers sets (yellow bars), the 100-fold potency difference threshold for AC formation was met. (b) Shown are exemplary isomer cliffs (with ΔpK_i values given in red) for two isomer sets formed by serotonin 6 (5-HT6) receptor ligands.

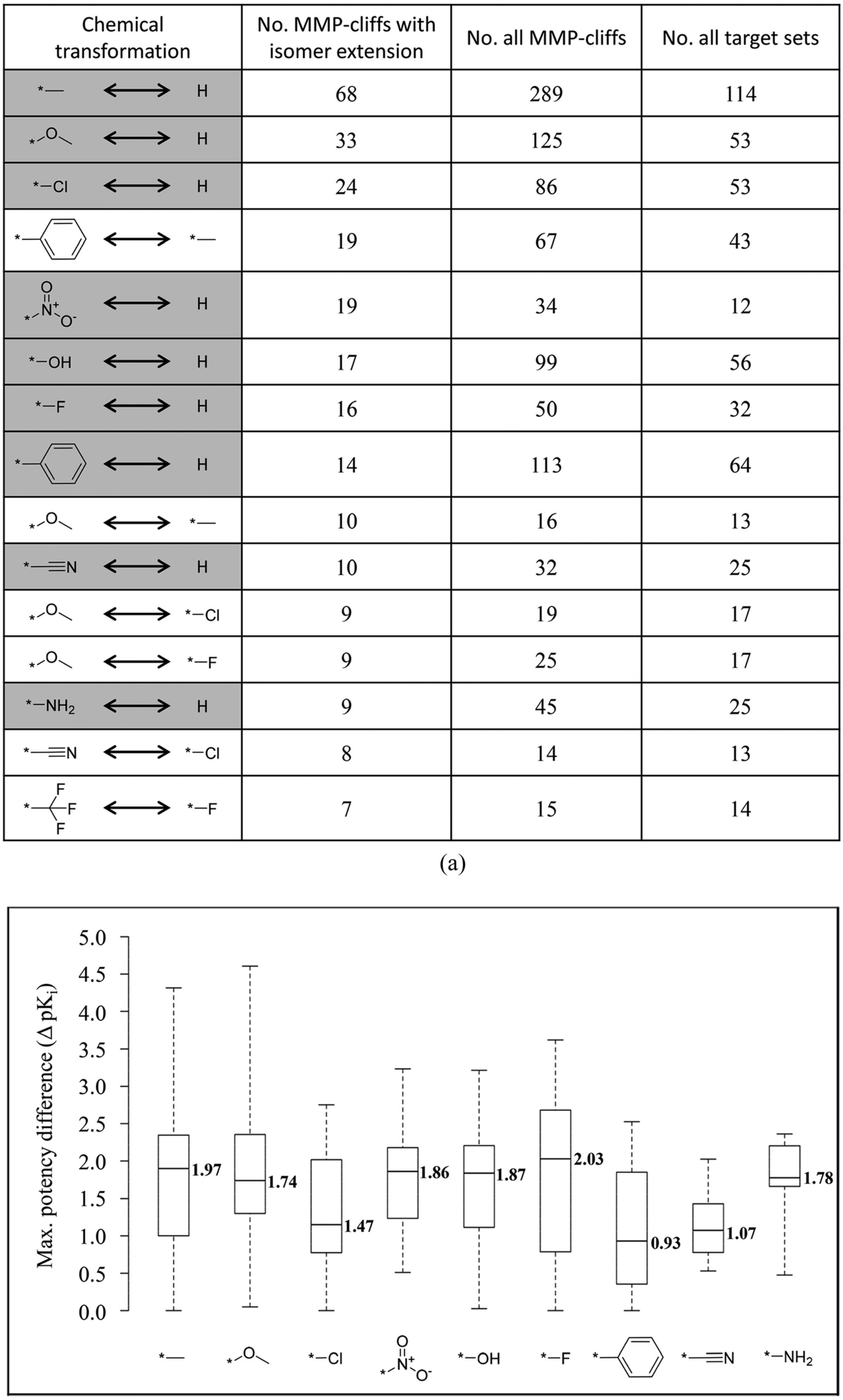
nicely illustrate "magic methyl" effects as a consequence of positional variations. Fig. 6b depicts corresponding examples of extended MMP-cliffs with an H ↔ F transformation, which revealed potency effects of fluorine substitutions at varying positions.

Among the 1182 MMP-cliffs with isomer extension, structural isomers of weakly potent, highly potent, or both cliff compounds were detected for 589, 496 and 97 MMP-cliffs, respectively (with a median value of one isomer per MMP-cliff). Fig. 7 shows examples of MMP-cliffs with different chemical transformations for which structural isomers of both weakly and highly potent cliff compounds were available. These examples illustrate various effects of methyl to phenyl or chlorine to bromine replacements at different positions.

Taken together, the representative examples discussed above reveal that extension of MMP-cliffs with structural isomers was SAR-informative even in cases where the AC potency difference threshold was not reached and no formally defined isomer/MMP-cliffs were obtained. However, isomer extension generally resulted in an increase in relevant compound relationships and positional effects of substitutions provided additional SAR information for nearly 1200 MMP-cliffs with activity against 147 targets (Fig. 3).

Isomer/MMP-cliffs

Our analysis yielded a total of 597 isomer/MMP-cliffs, which represented the subset of MMP-cliffs with isomer extension having largest potency effects. Compared to MMP-cliffs, the small number of currently available isomer/MMP-cliffs indicates that SARs involving isomers of specific substitutions are only little explored. This also



Chemical transformation	No. MMP-cliffs with isomer extension	No. all MMP-cliffs	No. all target sets
*— ⟷ H	68	289	114
*—O— ⟷ H	33	125	53
*—Cl ⟷ H	24	86	53
*—(phenyl) ⟷ *—	19	67	43
*—N⁺(=O)O⁻ ⟷ H	19	34	12
*—OH ⟷ H	17	99	56
*—F ⟷ H	16	50	32
*—(phenyl) ⟷ H	14	113	64
*—O— ⟷ *—	10	16	13
*—≡N ⟷ H	10	32	25
*—O— ⟷ *—Cl	9	19	17
*—O— ⟷ *—F	9	25	17
*—NH_2 ⟷ H	9	45	25
*—≡N ⟷ *—Cl	8	14	13
*—CF_3 ⟷ *—F	7	15	14

(a)

(b)

Fig. 5 Transformations in MMP-cliffs with isomer extension. (a) Listed are the top 15 most frequent chemical transformations in MMP-cliffs with isomer extension. Nine transformations representing hydrogen atom replacements are highlighted using a gray background. (b) For these nine chemical transformations, the distribution of maximum pairwise potency differences between MMP-cliff compounds and structural isomers is shown in boxplots. In each case, the median value is reported.

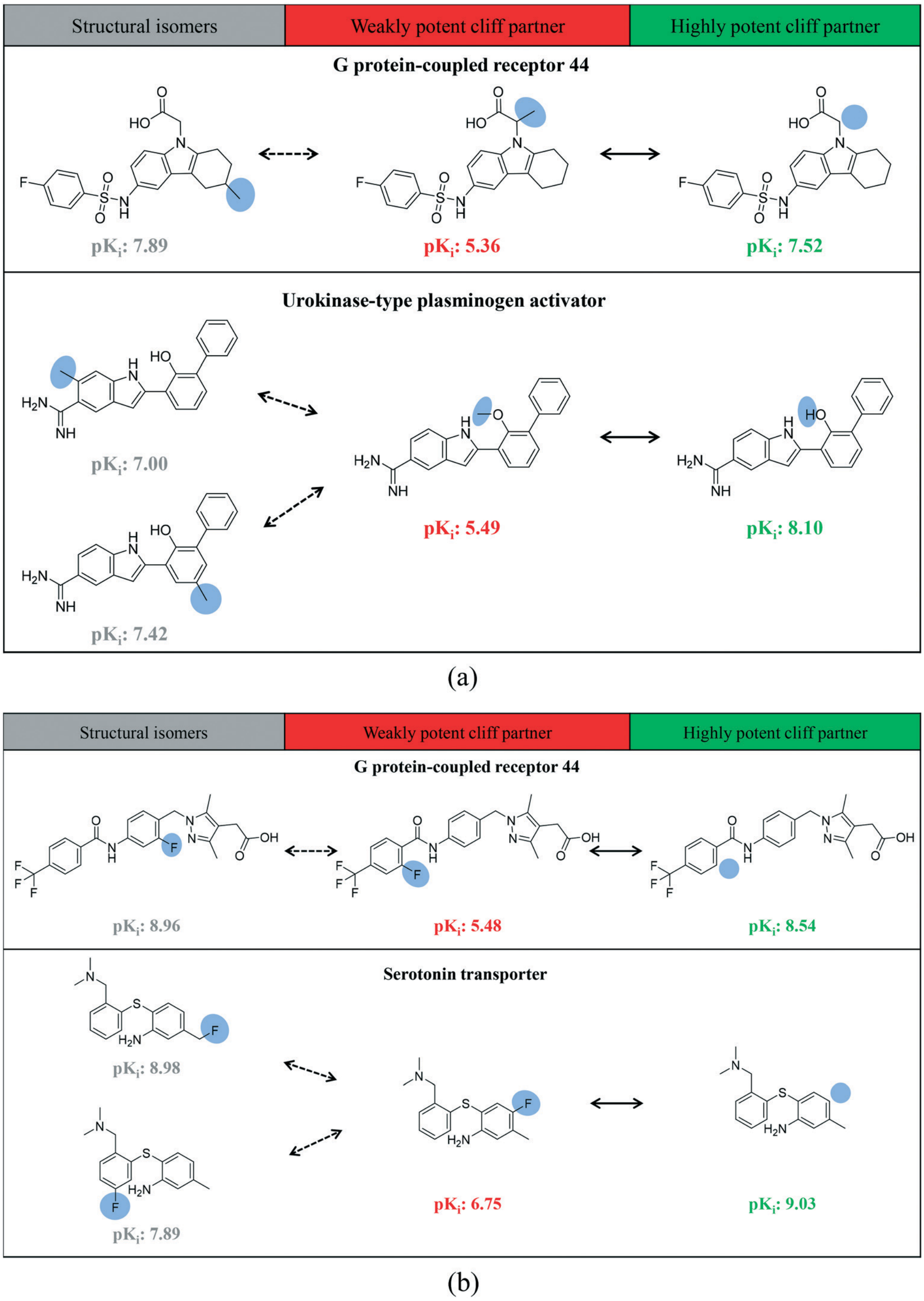


Fig. 6 MMP-cliffs with smallest transformations and isomer extension. Shown are exemplary MMP-cliffs with isomer extension that captured the smallest possible chemical transformation including the replacement of a hydrogen atom with a (a) methyl group (H ↔ CH_3) and (b) fluorine atom (H ↔ F). Compound targets are given.

indicates that newly introduced substituents at a given sites are only infrequently considered at other positions in active compounds. This has implications for practical medicinal chemistry and suggests further analog design strategies such as the introduction of new substituents at proximal yet distinct sites (yielding an MMP with isomer extension).

Fig. 8 shows exemplary isomer/MMP-cliffs for which qualifying structural isomers of highly potent (Fig. 8a) or weakly potent cliff compounds (Fig. 8b) were available. The examples illustrate how an isomer/MMP-cliff transforms a standard MMP-cliff into an AC with different substituents at two sites, thereby extending the MMP-formalism according to which substitutions are limited to

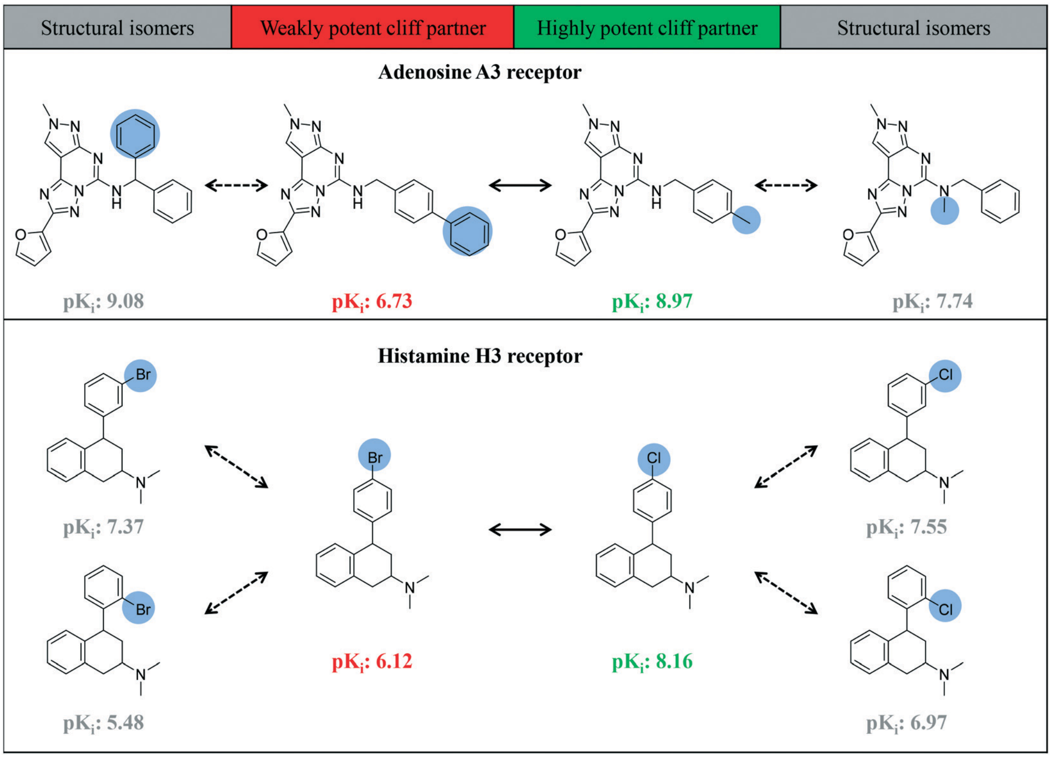


Fig. 7 Fully extended MMP-cliffs. Exemplary MMP-cliffs are shown for which structural isomers of both highly and weakly potent cliff compounds were available.

a single site. ACs with substitutions at multiple sites can also be obtained if they are extracted from analog series including computationally identified series.[9] In Fig. 8a, structural isomers of highly potent MMP-cliff compounds have comparable potency. In Fig. 8b, isomers of weakly potent cliff partners display larger potency variations, but the pre-defined AC potency difference threshold is met in both instances. Comparison of MMP-cliffs and corresponding isomer/MMP-cliffs makes it possible to better understand if the chemical nature of a substitution and/or its position might be more important for achieving high compound potency, which can be further assessed through the design of additional analogs. Since only a small proportion of isomer sets contained compounds with potency variations of large magnitude, as also shown herein, isomer/MMP-cliffs are also likely to indicate regions in potent compounds where key substituents might be positioned in different ways, as illustrated in Fig. 8, hence providing alternatives for chemical synthesis.

Conclusions

In this study, we have introduced a new category of ACs by associating MMP-cliffs with structural isomers, leading to the definition of isomer/MMP-cliffs. These ACs uniquely combine different similarity criteria and transform MMP-cliffs into ACs with different substituents at two sites. Through large-scale compound data analysis, the presence of isomer cliffs and isomer/MMP-cliffs in different target sets was confirmed and a data set of MMP-cliffs with isomer extension was obtained. We have shown that isomer extension provides additional SAR information for MMP-cliffs, regardless of whether isomer/MMP-cliffs are formed or not, which depends on the chosen potency difference threshold. Hence, MMP-cliffs with isomer extension and isomer/MMP-cliffs might be considered to reveal a continuum of SARs and potency effects, rather than as discrete states. Regardless, the newly introduced data structure is highly SAR-informative. In some instances, very small chemical modifications such as the introduction of a methyl group at varying positions led to significant potency alterations in active compounds. In others, positional variation of larger substituents that were critical for high potency was readily tolerated. Such findings make this data structure interesting for SAR exploration in medicinal chemistry. From a computational perspective, isomer/MMP-cliffs are also thought to provide meaningful test cases for potency prediction methods. Hence, taken together, the extension of the MMP-cliff concept and new AC category introduced herein widen the current spectrum of AC and provide additional opportunities for SAR exploration. These opportunities also include complementary analysis of new two- and three-dimensional ACs,[17] which can be extended through structure-based predictive modeling.[18] For SAR analysis or other investigations, our data set of MMP-cliffs with isomer extension is freely available upon request.

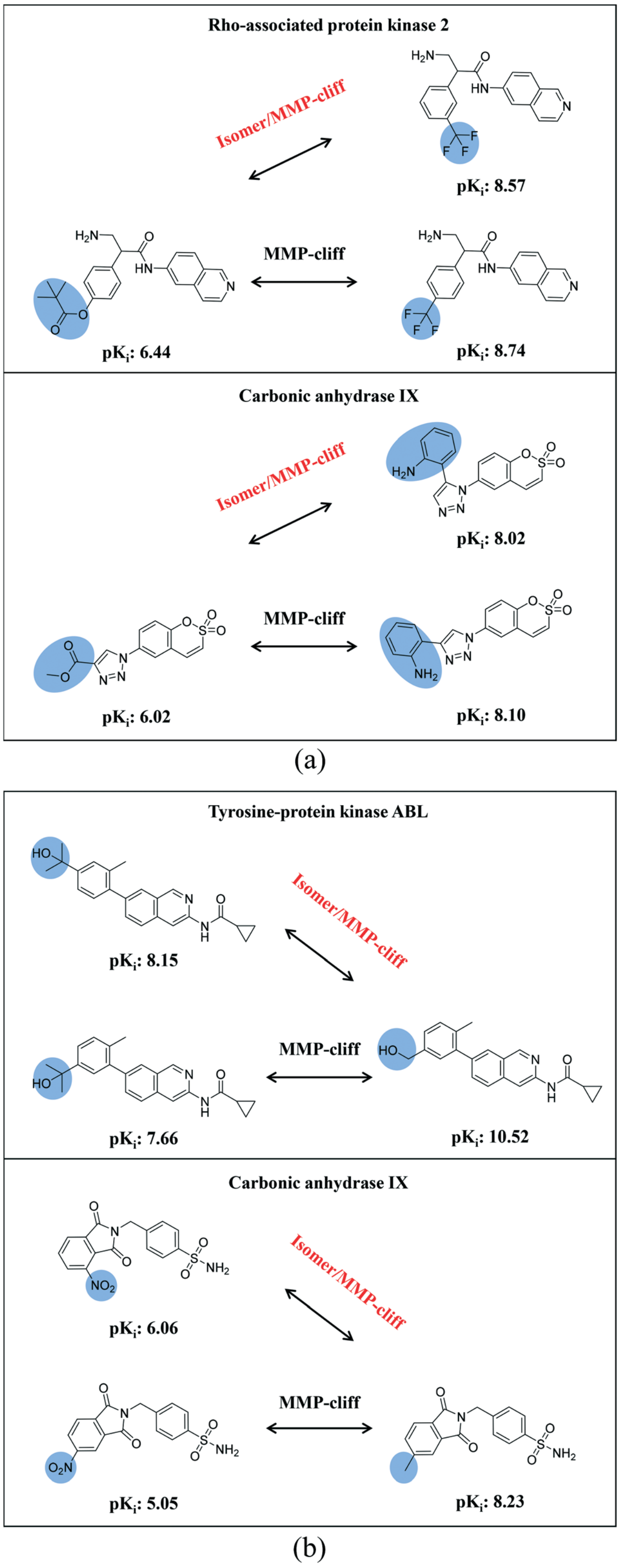


Fig. 8 Isomer/MMP-cliffs. Shown are exemplary isomer/MMP-cliffs where structural isomers replaced the (a) highly or (b) weakly potent MMP-cliff compounds.

Conflicts of interest

There is no conflict of interest to declare.

Acknowledgements

The authors thank Dr. Terry R. Stouch for helpful suggestions and encouragement and Dr. Dagmar Stumpfe for helpful discussions and critical review of the manuscript. Furthermore, we thank the OpenEye Free Academic Licensing Program for providing a free academic license for the chemistry toolkit. H. H. is supported by the China Scholarship Council.

References

1 G. M. Maggiora, *J. Chem. Inf. Model.*, 2004, **46**, 1535.

2 D. Stumpfe and J. Bajorath, *J. Med. Chem.*, 2012, **55**, 2932–2942.

3 D. Stumpfe, D. Y. Hu, Y. D. Dimova and J. Bajorath, *J. Med. Chem.*, 2014, **57**, 18–28.

4 G. M. Maggiora, M. Vogt, D. Stumpfe and J. Bajorath, *J. Med. Chem.*, 2014, **57**, 3186–3204.

5 Y. Hu and J. Bajorath, *J. Chem. Inf. Model.*, 2012, **52**, 1806–1811.

6 J. Hussain and C. Rea, *J. Chem. Inf. Model.*, 2010, **50**, 339–348.

7 X. Hu, Y. Hu, M. Vogt, D. Stumpfe and J. Bajorath, *J. Chem. Inf. Model.*, 2012, **52**, 1138–1145.

8 D. Stumpfe, D. Dimova and J. Bajorath, *J. Med. Chem.*, 2016, **59**, 7667–7676.

9 D. Stumpfe, H. Hu and J. Bajorath, *Bioorg. Med. Chem.*, 2019, **27**, 3605–3612.

10 Y. Hu, N. Furtmann and J. Bajorath, *RSC Adv.*, 2015, **5**, 43006–43015.

11 H. Hu, D. Stumpfe and J. Bajorath, *Future Med. Chem.*, 2019, **11**, 379–394.

12 Y. Hu, D. Stumpfe and J. Bajorath, *F1000Research*, 2013, **2**, e199.

13 D. Stumpfe and J. Bajorath, *Future Med. Chem.*, 2015, **7**, 1565–1579.

14 A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte, M. Davies, N. Dedman, A. Karlsson, M. P. Magariños, J. P. Overington, G. Papadatos, I. Smit and A. R. Leach, *Nucleic Acids Res.*, 2017, **45**, D945–D954.

15 *OEChem TK*, OpenEye Scientific Software, Inc., Santa Fe, NM, USA, 2012.

16 N. Schneider, R. A. Lewis, N. Fechner and P. Ertl, *ChemMedChem*, 2018, **13**, 1315–1324.

17 Y. Hu, N. Furtmann and J. Bajorath, *RSC Adv.*, 2015, **5**, 43006–43015.

18 J. Husby, G. Bottegoni, I. Kufareva, R. Abagyan and A. Cavalli, *J. Chem. Inf. Model.*, 2015, **55**, 1062–1076.

Summary

In this chapter, the AC concept was extended by combining the structural isomer and MMP concepts. Firstly, MMP-cliffs were generated using the ChEMBL database. For the sake of identifying structural isomers of MMP-cliff compounds, the attachment site of the core was replaced by hydrogen and structural isomers of the MMP compounds with the same core were identified. These isomers also had to display a potency difference of at least two orders of magnitude compared to the MMP partner in order to form an isomer/MMP-cliff. In general, only for a limited percentage of MMP-cliffs (4.4%), structural isomers were found implying that structural isomers are typically less studied in compound optimization. For MMP-cliffs with isomers, around 40% of MMP-cliffs could be extended with additional compounds with significantly large differences in potency, resulting in 597 new ACs. Although this AC category is less frequently observed than MMP-cliffs, it is richer in SAR information. Since a large proportion of MMP-cliffs could not be extended to isomer/MMP-cliffs, i.e., no structural isomers were available, the study of this AC category suggested to further consider positional alteration of R-groups in compound optimization.

In the next chapter, we relate the privileged substructure concept to activity cliffs and systematically explore the frequency of occurrence of privileged substructures in an AC analysis.

Chapter 4

Systematic Exploration of Activity Cliffs Containing Privileged Substructures

Introduction

The privileged substructure (PS) concept was originally introduced to identify core structures that preferentially occurred in compounds with activity against a given target family. Subsequently, many target family-privileged substructures have been proposed. However, an increasing number of studies indicate that their high prevalence in bioactive compounds does not necessarily imply that they are exclusively active against a desired target family.

In this chapter, we analyzed PS distributions among different target families, and for the first time, studied PSs in ACs. Different types of ACs were systematically extracted and further divided into two different AC categories depending on whether they contained predefined PSs or not. A comprehensive analysis in comparing ACs with and without PSs was conducted.

Reprinted with permission from “Hu, H.; Bajorath, J. Systematic Exploration of Activity Cliffs Containing Privileged Substructures. *Molecular Pharmaceutics* **2020**, *17*, 979-989”. *Copyright 2020 American Chemical Society*

molecular pharmaceutics

pubs.acs.org/molecularpharmaceutics Article

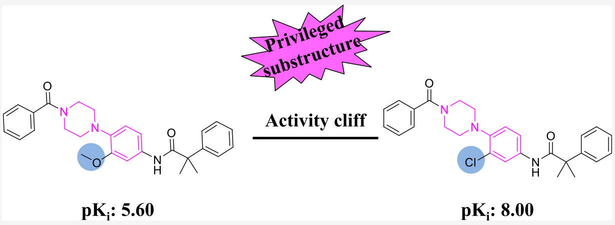
Systematic Exploration of Activity Cliffs Containing Privileged Substructures

Huabin Hu and Jürgen Bajorath*

Cite This: *Mol. Pharmaceutics* 2020, 17, 979−989 Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: The privileged substructure (PS) and activity cliff (AC) concepts are popular in pharmaceutical research. PSs have been empirically identified as preferred building blocks for target-class-directed generation of active compounds. Although some PSs are controversially viewed, they continue to receive much attention in drug discovery. ACs are formed by structurally similar active compounds with large potency differences and thus capture structure−activity relationship (SAR) discontinuity and reveal SAR determinants. So far, the PS and AC concepts have not been investigated in context. We have systematically explored ACs formed by compounds containing different PSs (PS-ACs). Such ACs were frequently identified in different series of compounds. PS-ACs were thoroughly characterized and compared to ACs formed by other compounds. The analysis revealed differences in AC formation between PSs. For example, individual PSs with an unusually high proportion of AC-forming compounds were identified. Furthermore, PS-AC network analysis identified clusters of ACs containing the same PS in different compound structure contexts with activity against different targets. From such PS-AC clusters, target-specific SAR information for PSs in different structural environments can be extracted.



KEYWORDS: *activity cliffs, privileged substructures, target classes, structure−activity relationships, compound design*

INTRODUCTION

The activity cliff (AC) concept is widely applied in computational medicinal chemistry and drug design.[1] Defined as pairs of structurally analogous active compounds with large differences in potency against their targets,[1] ACs represent the apex of structure−activity relationship (SAR) discontinuity.[1,2] Accordingly, they have high SAR information content and are focal points of SAR exploration in medicinal chemistry. During compound optimization, ACs are typically considered on a case-by-case basis, without the need to formally specify cliff criteria. By contrast, for computational identification and consistent evaluation of ACs—including investigations of large compound data sets—similarity and potency difference criteria for AC formation must be unambiguously defined and consistently applied.[1,3]

The AC concept has been evolved and refined over time.[2−5] Alternative molecular similarity approaches have been investigated including fingerprint-descriptor- and substructure-based similarity.[3−5] The use of descriptor-based numerical similarity values requires the application of thresholds for AC formation, which are typically subjectively defined. Alternatively, substructure-based similarity criteria are applicable such as the presence of a conserved compound core structure (representing a series of analogues)[3,4] or the formation of a matched molecular pair (MMP),[6−8] defined as a pair of compounds that are only distinguished by a chemical change at a single site.[6] Core-structure- or analogue-series-based similarity assessment makes it possible to consider multiple substitution sites for AC formation.[5] Furthermore, potency distributions of active compounds have been evaluated on a large scale and in an activity-class-dependent manner in order to determine the significance of potency differences between structural analogues for AC formation.[9,10] Taken together, these investigations have led to the introduction of alternative definitions and different types of ACs for a variety of applications.[4,5]

The privileged substructure (PS) concept is also very popular in medicinal chemistry.[11−14] It is not related to the AC concept. Originally introduced by Evans et al.,[11] PSs represent recurrent substructures in compounds with preferential activity against specific target families or classes,[11,12] Although PSs are typically not specific for a given target class,[15] their tendency to preferentially bind to, for example, G protein coupled receptors (GPCRs)[16] or kinases[17,18] continues to be of high interest for target-class-directed compound design.[12,13] Typical examples include the benzodiazepine substructure found in many GPCR

Received: December 6, 2019
Revised: January 22, 2020
Accepted: January 24, 2020
Published: January 24, 2020

https://dx.doi.org/10.1021/acs.molpharmaceut.9b01236

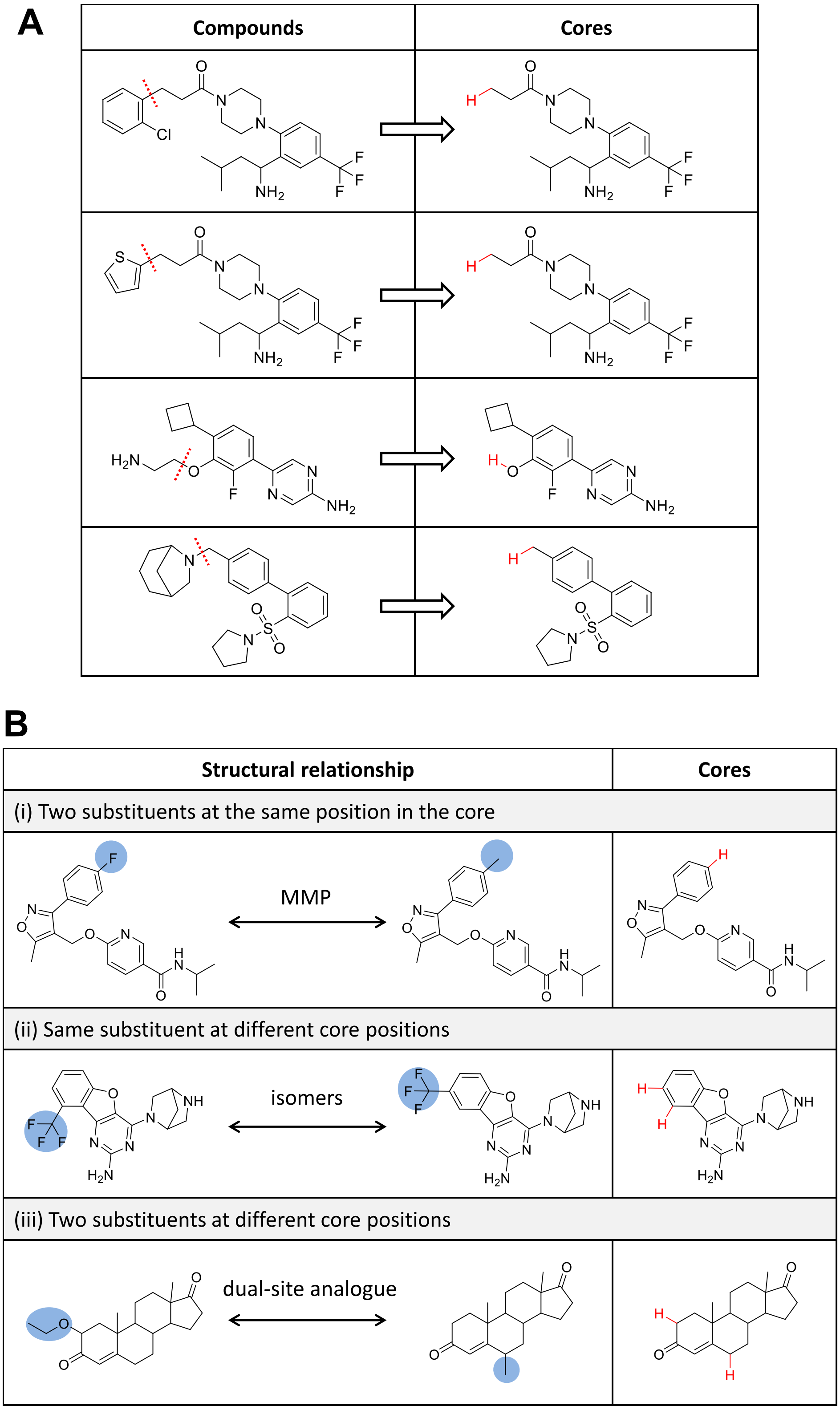

A
Compounds
Cores
B
Structural relationship
Cores
(i) Two substituents at the same position in the core
MMP
(ii) Same substituent at different core positions
isomers
(iii) Two substituents at different core positions
dual-site analogue

Figure 1. Compound core generation and structural relationships. (a) For four exemplary compounds, the construction of an exemplary core is illustrated. The generation involves single-site cleavage (red dashed line) of a bond to a qualifying substituent, followed by hydrogen atom replacement (red) in the core. (b) For compounds sharing a given core, three types of structural relationships are illustrated. Substituents distinguishing paired compounds are highlighted using a blue background.

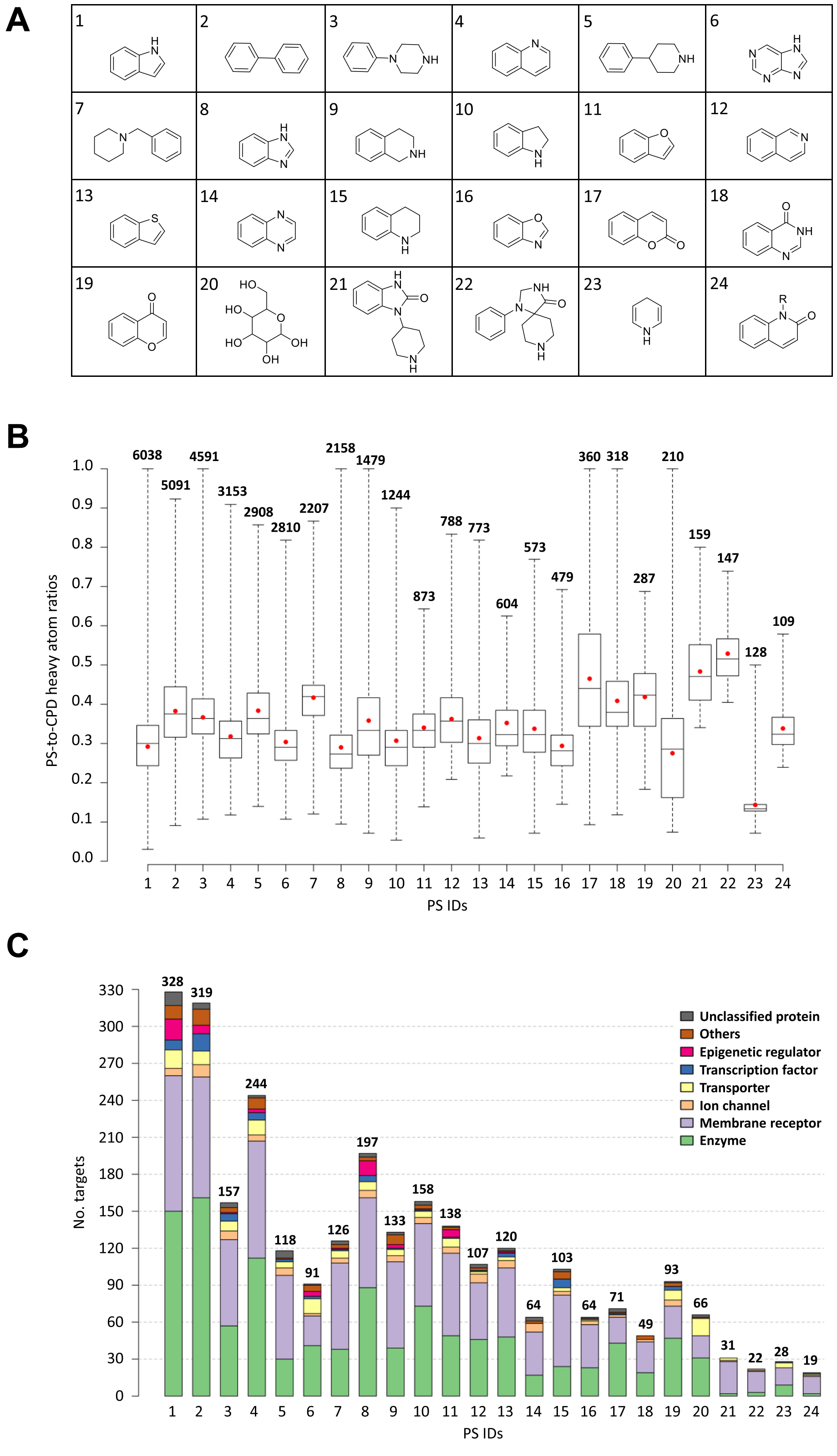


Figure 2. Privileged substructure analysis. (a) Shown are 24 PSs. The given numerical PS identifiers (IDs) are consistently used. (b) For each PS, the distribution of PS-to-compound (CPD) heavy atom ratios is reported in a boxplot. Red dots indicate mean values. The total number of active compounds containing each PS is given above each plot. (c) The bar graph reports the distribution of PSs over targets (color-coded according to classes). The total number of targets is given above each bar. PSs are arranged in the order of decreasing compound numbers.

ligands[11,16] or the quinazoline moiety that is recurrent in kinase inhibitors.[18] Origins of PS characteristics at the molecular level of detail are yet to be fully explored. However, their high frequency of occurrence in ligands active against different target classes indicates that PSs typically yield compounds with sustainable SARs for these targets.

In this work, we have investigated the AC and PS concepts in context by systematically searching for different types of ACs formed by compounds containing PSs. These ACs were analyzed and compared to others without PSs. AC network analysis distinguished between PSs with different SAR characteristics.

■ METHODS AND MATERIALS

Compounds and Activity Data. From ChEMBL[19] (version 25), compounds forming direct interactions with a single human target protein (i.e., assay relationship type "D") at the highest confidence level (i.e., assay confidence score 9) were extracted. As potency measurements, numeric-assay-independent equilibrium constants (K_i values) with standard relationship "=" were required. If a compound had multiple K_i values for the same target, the geometric mean of all values was calculated as the final potency, provided all values fell within the same order of magnitude (otherwise, the compound was disregarded). On the basis of these selection criteria, 77 189 unique compounds with activity against 962 targets were obtained (yielding a total of 130 810 activity annotations).

Target Classes. Biological targets of active compounds were assigned to eight classes following the ChEMBL target-classification scheme.[19] These included six major target classes (*enzymes*, *membrane receptors*, *ion channels*, *transporters*, *transcription factors*, and *epigenetic regulators*). In addition, *others* and *unclassified* referred to the union of several small target classes and targets not assigned to other classes, respectively.

Compound Core Generation. Applying a fragmentation algorithm for MMP generation,[7] single exocyclic bonds in compounds were systematically cleaved yielding two fragments per operation. Fragments were accepted on the basis of the following size restrictions:[8] The number of non-hydrogen atoms of one fragment was required to be at least twice as large as that of the other, and the size of the smaller fragment was confined to at most 13 non-hydrogen atoms. The larger fragment was defined as the "core", and the smaller was defined as the "substituent". In the core, the cleaved off smaller fragment was replaced with a hydrogen atom. Core generation is illustrated in Figure 1a. The calculations were carried out using in-house scripts with the aid of OpenEye chemistry toolkit.[20]

Structural Relationships. For all generated cores, compound pairs sharing the same core were systematically identified. Accordingly, each pair combined two structural analogues. Paired compounds were then computationally examined for the presence of three different types of structural relationships, as illustrated in Figure 1b. Specifically, compounds forming a pair were distinguished by

(i) *Two substituents at the same position in the core.* This relationship corresponded to the formation of an MMP.

(ii) *The same substituent at different core positions.* This relationship represented a pair of structural isomers.

(iii) *Two substituents at different core positions.* This relationship captured a pair of analogues with substitutions at two sites, termed dual-site analogues (ds-analogues). For ds-analogues, an additional size restriction was introduced. Accordingly, substituents at both sites were permitted to differ by at most eight non-hydrogen atoms.

Activity Cliffs. As a constant potency difference threshold for AC formation, an at least 100-fold difference between K_i values of structural analogues forming a qualifying pair was required. On the basis of the different pairwise structural relationships specified in Figure 1b, three types of ACs were identified:

(i) *MMP-cliffs (MMP-ACs).*[8] Paired analogues differ by substitutions at a single site.

(ii) *Isomer-cliffs (iso-ACs).* Paired analogues are structural isomers.

(iii) *Dual-site-ACs (ds-ACs).*[21] Paired analogues differ by substitutions at two sites.

Privileged Substructures. The PS collection of Welsch et al.[14] was used, which mostly originated from drugs and natural products. PSs were selected if they were found in at least 100 bioactive compounds by substructure searching. A given compound might contain more than one PS. Figure 2a shows 24 PSs that qualified for our analysis.

Compound Properties. For each AC compound, the fraction of sp^3 carbon atoms, topological polar surface area, and log*P* (octanol/water partition coefficient) were calculated with the RDKit toolkit[22] implemented in KNIME protocols.[23]

In addition, ligand efficiency (LE) was computed using the binding efficiency index (BEI)[24] defined as

$$\mathrm{LE} = \frac{\mathrm{p}K_i}{\mathrm{MW}}[\text{log unit/kDa}]$$

, in which MW stands for molecular weight. Since LE were only compared for structural analogues, corrections for potential size dependence[25] were not required.

■ RESULTS AND DISCUSSION

Considering Privileged Substructures and Activity Cliffs in Context. In our study, we have aimed to identify ACs involving PSs and compare these ACs with others formed by compounds without PSs. ACs with PSs have thus far not been considered. In pharmaceutical research, target-class preferences of PSs were for the most part empirically assigned on the basis of expert knowledge and compounds becoming available over time. While there is no general rationale for target-class selectivity of PSs, preferential binding has often been attributed to class-characteristic binding patterns,[12,16] resulting in a tendency to enrich compounds containing PSs with certain bioactivities. As such, PSs are attractive for compound design and SAR exploration and hence fall into the applicability domain of the AC formalism. ACs capture SAR discontinuity among active compounds and indicate optimization potential of candidate compounds. We investigated the formation of ACs among PS-containing compounds and also compared PSs on the basis of AC analysis, thereby establishing a link between the PS and AC concepts.

Privileged Substructure Analysis. Initially, we analyzed the 24 PSs in Figure 2a that were found in at least 100 active compounds in more detail. The PSs are indexed in the order of decreasing compound numbers. The majority of these PSs consisted of a condensed aromatic ring system with different heteroatom substitutions. According to current standards, planar aromatic compounds are not necessarily among preferred starting points for medicinal chemistry efforts. However, these PSs were recurrent in active compounds and prioritized over time, having a long history in medicinal chemistry.

The 24 PSs were detected in 31 773 (41.2%) of all active compounds. Hence, they were widely distributed, and the generated data set was suitable for our analysis. The most frequently occurring PS was indole (PS 1, 6038 compounds), followed by biphenyl (PS 2, 5091), arylpiperazine (PS 3, 4591), and quinoline (PS 4, 3153). Figure 2b reveals that active compounds containing PSs were typically more than twice as large as the respective PS, which likely resulted from compound optimization efforts. We also examined the target distribution of PS compounds on the basis of eight broadly defined target classes. The membrane receptor class was mostly comprised of GPCRs, while the enzyme class contained many different families that were not further differentiated. As shown in Figure 2c, this general target classification was sufficient to confirm that compounds containing PSs were not class-specific, consistent with earlier findings. While there was a tendency for most PSs to occur in compounds active against membrane receptors and various enzymes, which dominated the distribution, all PSs were found in compounds with activity against different target classes.

Identification of Activity Cliffs. A systematic search for three types of ACs was carried out including MMP-ACs, iso-ACs, and ds-ACs. As described in Methods and Materials, these types of ACs were distinguished by different structural relationships between participating analogues including substitutions at a given site in a conserved core structure (MMP-ACs), substitutions at two sites (ds-ACs), or topology differences (iso-ACs).

The systematic search for ACs is outlined in Figure 3, and the results are summarized. A total of 704 019 compound pairs meeting AC similarity criteria were identified that were active against 677 targets. Applying a ΔpK_i ≥ 2.0 potency difference threshold, a total of 34 049 ACs were obtained that involved 16 821 unique compounds with activity against 373 targets. These ACs included 26 584 MMP-ACs, 6945 ds-ACs, and 520 iso-ACs. Thus, MMP-ACs represented the majority of cliffs. We then examined ACs for the presence of PSs and identified 15 919 ACs in which both compounds containing the same PS (PS-ACs). These ACs were formed by 6927 unique compounds with activity against 204 targets. In addition, 18 130 ACs were identified that did not contain PSs. These ACs involved 10 016 unique compounds that were active against 328 targets. Hence, comparably large numbers of PS-ACs and other ACs were obtained for further analysis. Overall, ~22% of all active compounds were involved in the formation of one or more ACs, consistent with earlier results of global AC analysis.[9]

Activity Cliffs with Privileged Substructures. Figure 4 shows exemplary PS-ACs. In Figure 4a, MMP-ACs for different receptors are depicted, and in Figure 4b, MMP-ACs, iso-ACs, and ds-ACs with activity against other receptors, carbonic anhydrase II, or the PI3-kinase p110-alpha subunit are depicted.

Table 1 details the distribution of ACs over PSs and their target coverage (PSs are numbered according to Figure 2a). The distribution was uneven, as anticipated on the basis of significantly varying numbers of compounds containing each PS (Figure 2b). For several PSs with low compound numbers, less than 50 ACs were identified. However, there were notable exceptions. For example, 159 compounds containing piperidinyl-benzimidazolone (PS 21) were available that formed 186 MMP-ACs. Furthermore, 360 compounds containing coumarin (PS 17) yielded 41 MMP-ACs, 5 iso-ACs, and 73 ds-ACs. However, although coumarin has a long history in medicinal chemistry and is classified as a PS, coumarin derivatives have also been implicated in assay interference effects.[26] Hence, a word of caution is advised when investigating coumarin as a PS. Similar concerns have also been raised about other well-recognized PSs such as chromone (PS 19), which might be reactive under certain assay conditions.[27] On the other hand, chromone is also found as a substructure in several marketed drugs.[14] Hence, it remains difficult to generalize undesirable assay interference effects,[28] especially for PSs.

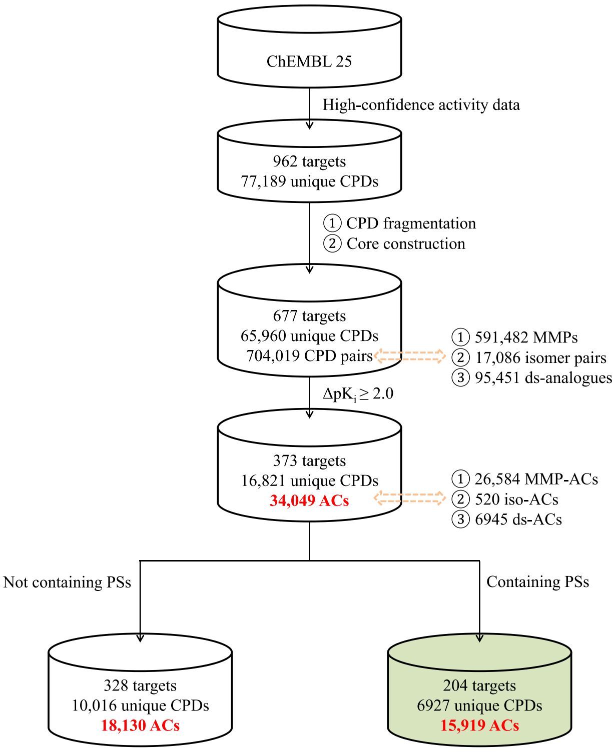


Figure 3. Identification of activity cliffs. The flowchart summarizes the identification of different types of ACs formed by compounds with and without PSs. For each analysis step, applicable compound, target, and AC statistics are provided.

Other PSs with large numbers of available compounds dominated the distribution of PS-ACs. There were 7 PSs for which more than 1000 ACs were identified. Biphenyl (PS 2) yielded by far the largest number of PS-ACs, with 5607 instances including 5094 MMP-ACs. Purine (PS 6) followed with 2109 instances (1218 MMP-ACs, 870 ds-ACs, and 21 iso-ACs), and arylpiperidine (PS 5) with 1482 ACs (including 1018 MMP-ACs). By far the largest target coverage was observed for indole- (78 targets) and biphenyl-containing ACs (72).

Taken together, the results showed that PS-ACs were consistently detected and that their frequency of occurrence roughly scaled with available compound numbers, with some exceptions. As observed globally, MMP-ACs also dominated the distribution of PS-ACs, whereas iso-ACs were only infrequently detected (consistent with the limited formation

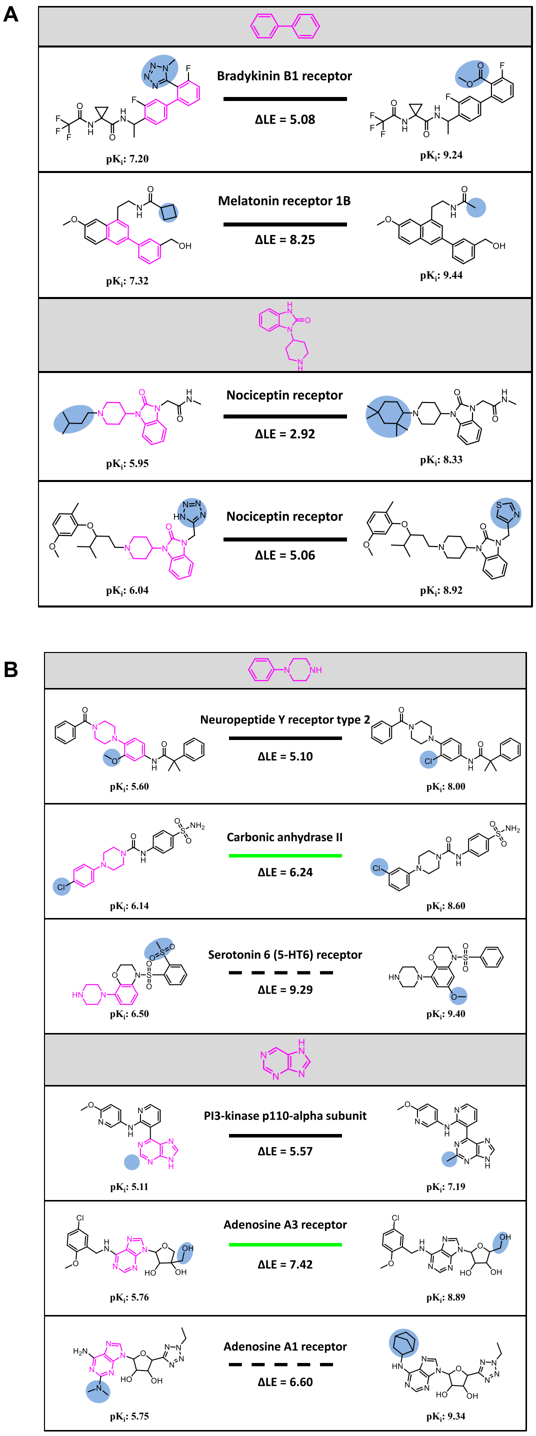


Figure 4. Exemplary activity cliffs with different privileged substructures. For selected PSs (depicted at the top in pink on a gray background), exemplary ACs of different types are shown. In each case, styles of lines connecting weakly (left) and highly potent (right) AC compounds indicate the cliff type (i.e., solid black line, MMP-AC; solid green, iso-AC; dashed black, ds-AC). In AC compounds on the left, the PS is colored pink. Distinguishing substituents are highlighted using a blue background. For each compound, the pK_i value is given. For each AC, the target is provided, and the LE difference between AC partners is reported. For different PSs, (a) shows MMP-ACs, and (b) shows ACs of all three types.

Table 1. Distribution of Activity Cliffs across Privileged Substructures[a]

PS IDs	total no. ACs	no. MMP-ACs	no. iso-ACs	no. ds-ACs
1	1298 (78)	844 (69)	37 (18)	417 (43)
2	5607 (72)	5094 (67)	36 (24)	477 (41)
3	1058 (31)	652 (30)	36 (9)	370 (19)
4	788 (44)	585 (40)	28 (13)	175 (23)
5	1482 (22)	1018 (20)	22 (10)	442 (14)
6	2109 (24)	1218 (23)	21 (6)	870 (9)
7	1206 (24)	725 (23)	27 (9)	454 (16)
8	687 (35)	572 (33)	7 (7)	108 (15)
9	465 (30)	310 (28)	14 (7)	141 (16)
10	181 (26)	127 (24)	5 (5)	49 (9)
11	351 (12)	170 (11)	9 (3)	172 (4)
12	1247 (10)	1157 (10)	4 (1)	86 (4)
13	105 (17)	87 (16)	5 (4)	13 (5)
14	339 (13)	234 (13)	0 (0)	105 (5)
15	81 (20)	49 (18)	1 (1)	31 (7)
16	48 (10)	34 (10)	1 (1)	13 (3)
17	119 (9)	41 (8)	5 (2)	73 (7)
18	53 (9)	27 (7)	4 (3)	22 (6)
19	44 (9)	24 (8)	3 (2)	17 (4)
20	5 (3)	5 (3)	0 (0)	0 (0)
21	186 (5)	186 (5)	0 (0)	0 (0)
22	85 (8)	79 (8)	2 (2)	4 (3)
23	14 (4)	7 (4)	0 (0)	7 (3)
24	27 (3)	27 (3)	0 (0)	0 (0)

[a] Reported are the numbers of three different types of ACs across PSs. For each PS, the number in parentheses gives the total number of targets of its ACs.

of analogue pairs by structural isomers). AC frequency was comparable for compounds with and without PSs. On average, an AC compound without PS was involved in 3.6 ACs, while a PS-containing AC compound participated in 4.6 ACs. Thus, compounds from PS-ACs displayed a tendency to form more AC relationships than others.

Property Analysis. We next analyzed molecular properties of ACs. First, ligand efficiency (LE) was investigated. Formation of ACs is generally associated with an increase in LE from the weakly potent to the highly potent cliff compound. This is the case because participating structural analogues have comparable (or the same) size but a large difference in potency.

Figure 5 shows the distribution of ΔLE values for different types of ACs combining PS-ACs and others. On average, MMP-ACs were found to have the lowest ΔLE value (5.67). For iso-ACs and ds-ACs, mean ΔLE values were larger but comparable (7.22 vs 7.23). In the case of iso-ACs, the ΔLE value was solely a consequence of potency differences.

Next, ΔLE and Δlog*P* values were compared for PS-ACs and ACs without PSs, as shown in Figure 6a. The value distributions were very similar, with essentially no differences. Notably, AC formation was overall not accompanied by an increase in hydrophobicity from the weakly to the highly potent cliff partner (as is often observed during compound optimization). For highly potent cliff compounds extracted from ACs with and without PSs, LE, log*P*, and additional properties were analyzed including the fraction of sp^3 carbon atoms and topological polar surface area. The results are shown in Figure 6b. The only systematic difference was the lower polar surface area of highly potent cliff compounds from PS-

https://dx.doi.org/10.1021/acs.molpharmaceut.9b01236

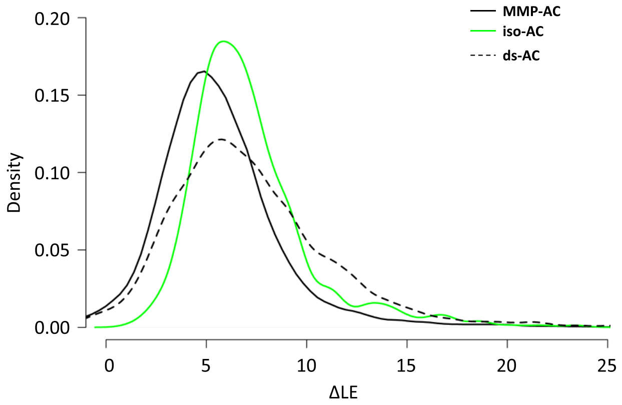


Figure 5. Ligand efficiency for activity cliffs. For different types of ACs with or without PSs (comprising all 34 049 ACs according to Figure 3), the density plot reports the distribution of LE differences between highly and weakly potent cliff partners ($LE_{highly} - LE_{weakly}$).

ACs (median value of 73.1 Å^2) than from ACs without PSs (median of 83.6 Å^2). However, although PS compounds mostly contained aromatic ring systems, the fraction of sp^3 carbon atoms of PS-AC and non-PS-AC compounds was similar (with a median value of 0.3). Taken together, the results showed that no significant molecular property differences were observed for PS-ACs compared to ACs without PSs.

Differences between Privileged Substructures. We further investigated the composition of PS-ACs. For each PS, the ratio of AC compounds to all compounds containing the PS was determined (Figure 7a, top) as well as the ratio of highly potent AC compounds to all AC compounds (bottom). The average ratio values of 21.8% for AC compounds to all PS compounds and of 57.3% for highly potent AC compounds to all AC compounds were very similar to corresponding mean ratios for ACs without PSs. However, PSs displayed a tendency to yield more highly than weakly potent AC compounds,

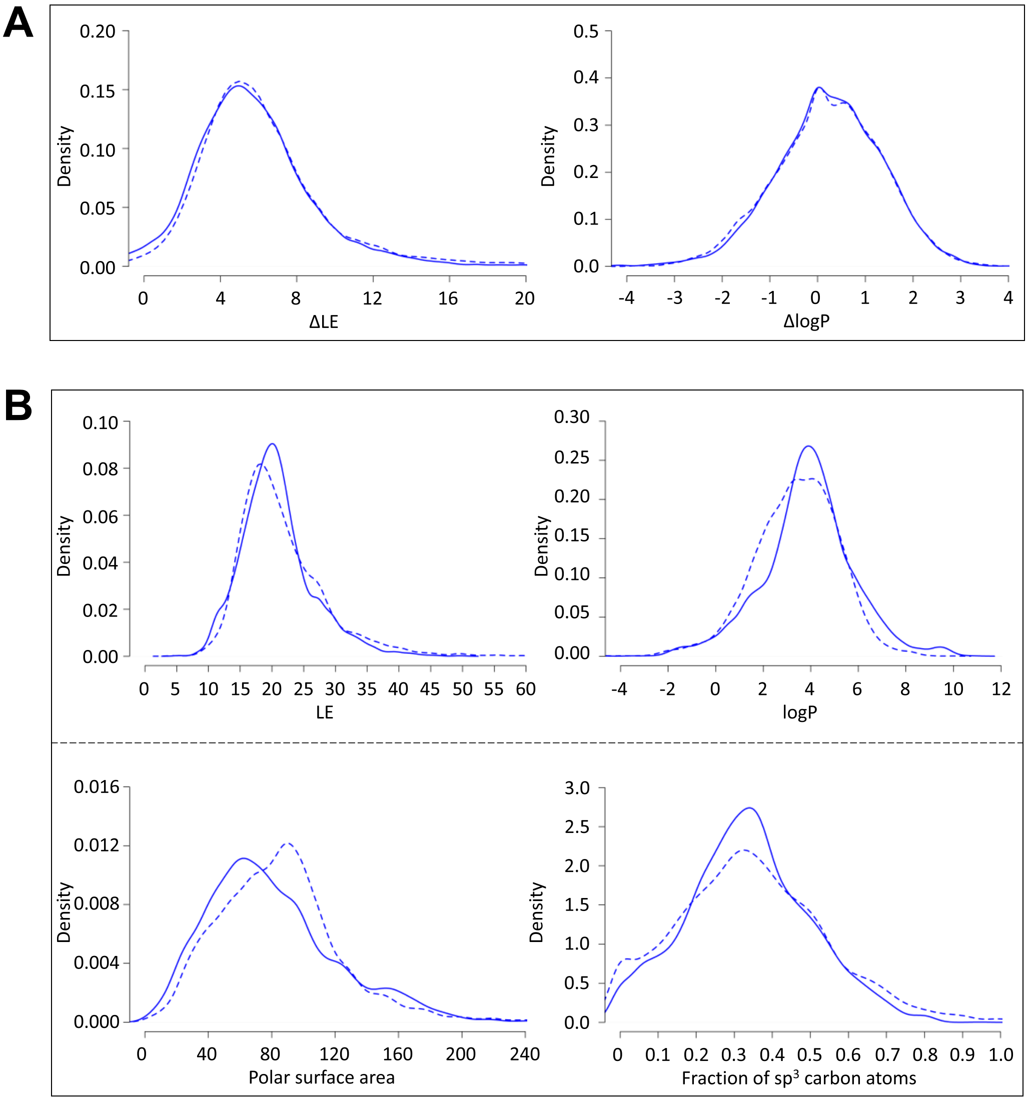


Figure 6. Comparison of molecular properties. Density plots compare property values for ACs with (solid line) or without (dashed line) PSs. (a) LE and log*P* differences between highly and weakly potent AC compounds are reported. (b) For highly potent AC compounds, values of different properties are compared.

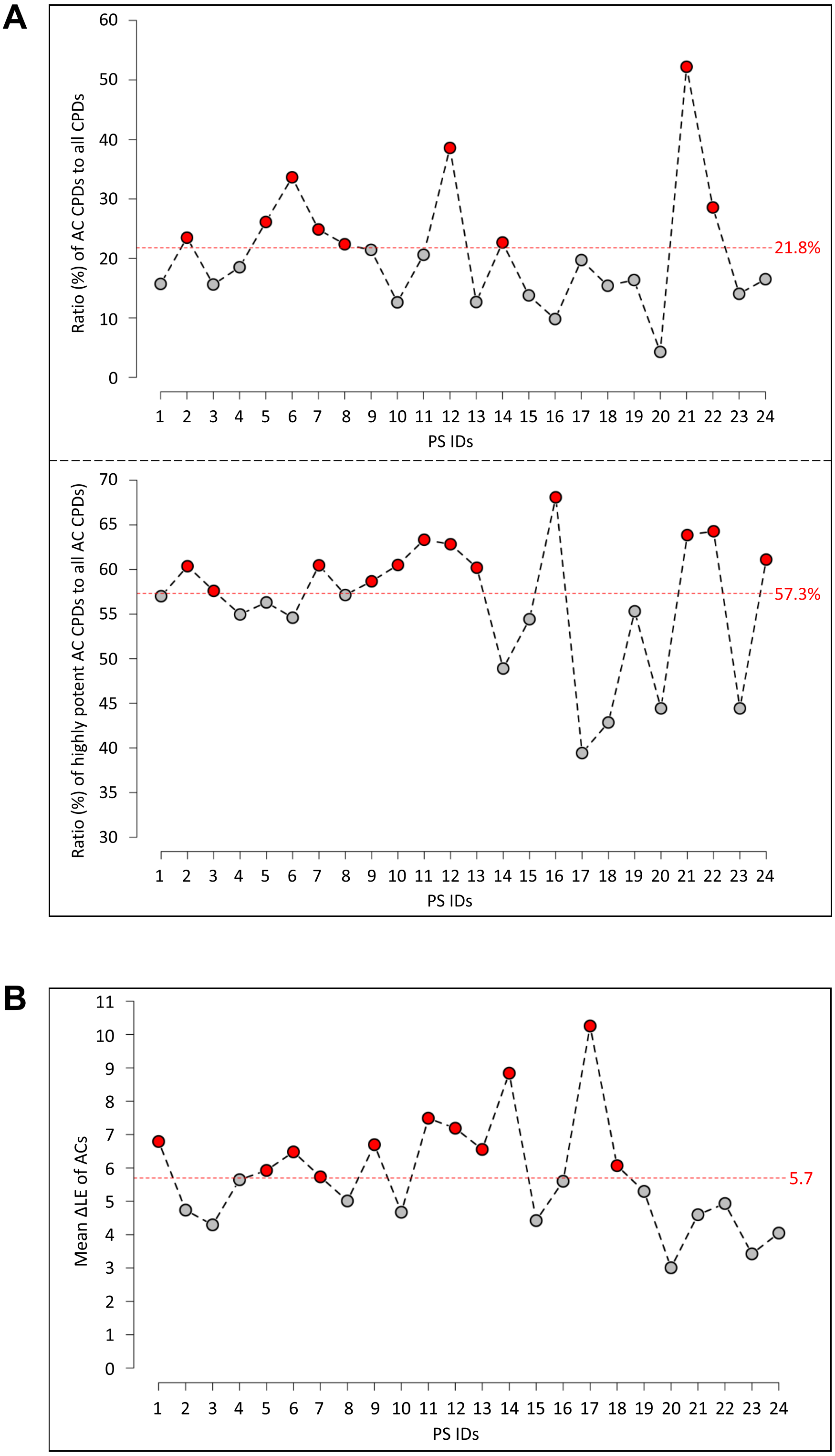


Figure 7. Compound proportions forming activity cliffs. For compounds (CPDs) containing each PS, (a) reports the ratio (%) of AC CPDs to all CPDs (top) and highly potent AC CPDs to all AC CPDs (bottom). In (b), the mean LE difference for ACs ($\text{LE}_{\text{highly}} - \text{LE}_{\text{weakly}}$) containing each PS is reported. Dashed red lines indicate global average values. For each PS, values above and below the global mean are marked with red and gray dots, respectively.

which was the case for 19 of 24 PSs (Figure 7a, bottom). In addition, the ratio strongly varied for several PSs with comparable compound numbers. For example, for benzoxazol (PS 16), nearly 70% of all AC compounds were highly potent. By contrast, for coumarin (PS 17), only 39% were highly potent.

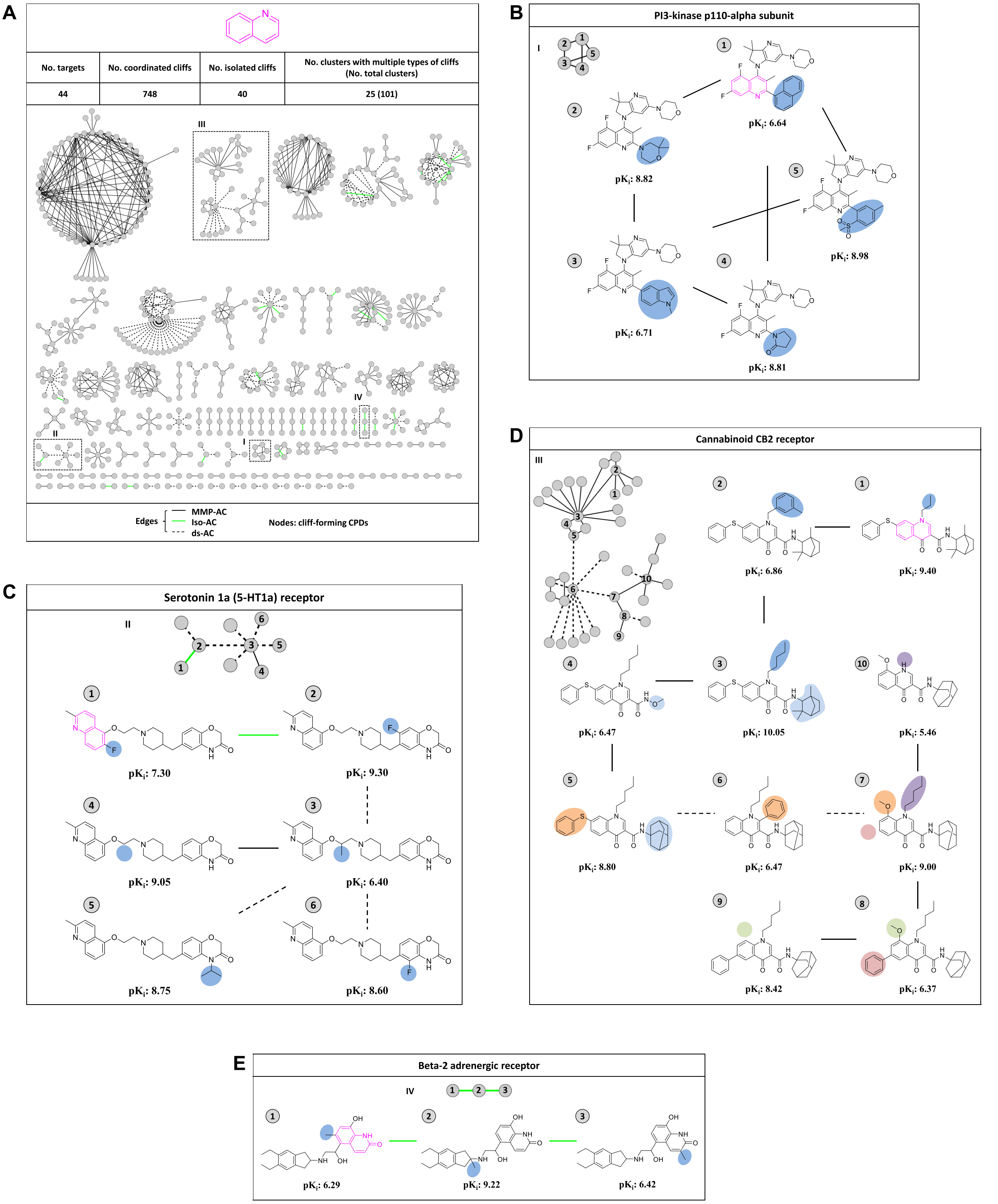


Figure 8. Activity cliff network analysis. (a) Shows the AC network representation for PS 4 (ID according to Figure 2a). In the network, nodes represent AC compounds and edges different types of pairwise ACs (indicated by line styles according to Figure 5). Network statistics are provided. Isolated ACs are formed by individual compound pairs, while coordinated ACs (with partners involved in multiple cliffs) give rise to the formation of increasingly large AC clusters. Four selected AC clusters are shown in dashed rectangles and identified using numbers I − IV. In (b−e), these clusters and their AC constellations are depicted in detail (the AC representation is according to Figure 5). AC compounds are numbered. In (d), substitutions distinguishing different ACs within the large cluster III are highlighted using differently colored backgrounds.

Further differences between PSs were observed for the ratio of AC compounds to all compounds. In this case, three PSs had a significantly higher proportion of AC compounds than others including purine (PS 6), isoquinoline (PS 12), and piperidinyl-benzimidazolone (PS 21), for which significantly different numbers of compounds were available. For PS 21, more than 50% of available compounds were involved in AC formation, an unusual observation. These findings indicated the presence of high SAR information content in sets of compounds with different PSs.

We also determined the mean ΔLE value accompanying AC formation for each PS, which revealed large differences between PSs (with large fluctuations around the mean), as shown in Figure 7b. For example, AC formation for PSs such as quinoxaline (PS 14) or coumarin (PS 17) was associated with unusually large LE improvements (with mean ΔLE = 8.8 and 10.3, respectively), whereas in other cases such as a carbohydrate substructure (PS 20), only small changes were observed (ΔLE = 3). Taken together, there were substantial differences in AC formation between PSs.

Activity Cliff Network Analysis. SAR information can be visualized in and extracted from AC networks, in which nodes represent AC compounds and edges indicate pairwise AC formation.[4,5] We also studied AC networks for PSs. Figure 8a shows an exemplary network for quinoline (PS 4) combining different types of ACs. Quinoline was contained in 3153 compounds, representing one of the largest PS-based sets. These compounds formed a total of 788 PS-ACs with activity against 44 targets (including, among others, 25 membrane receptors and nine enzymes). Only 40 ACs were formed by isolated analogue pairs. The others represented coordinated ACs where one or both AC compounds participated in multiple cliffs. In network representations, coordinated ACs lead to the formation of clusters (disjoint network components) with at least three compounds. These clusters typically have increasing size, different composition, and strongly varying topology. The 788 ACs depicted in Figure 8a formed a total of 101 clusters, 25 of which contained different types of ACs. The 748 coordinated ACs gave rise to 61 increasingly large clusters. Four representative clusters with different size and topology are marked in Figure 8a (I−IV) and their AC configuration is depicted in detail in Figure 8b−d, illustrating that AC clusters are particularly rich in SAR information.

Compounds forming cluster I in Figure 8b were PI3-kinase p110-alpha subunit inhibitors and only distinguished by substitutions at a single site (hence forming MMP-ACs). Different from cluster I, cluster II in Figure 8c contained all three types of ACs. Here, fluorine isomers alone displayed significant potency variations, which were complemented by other AC-inducing substitutions. We note that for structurally well-explored targets such as kinases, X-ray data might also be taken into consideration to further explore AC formation in the presence or absence of PSs.[5]

All compounds forming cluster III in Figure 8d were active against the cannabinoid CB2 receptor. This large cluster with complex topology contained a variety of coordinated ACs and was particularly rich in SAR information. Exemplary ACs are shown in detail revealing different substitution patterns with large potency effects. By contrast, the small cluster IV in Figure 8e contained only three compounds with activity against the beta-2 adrenergic receptor. The analogues formed iso-ACs involving a methyl group, which revealed a "magic methyl" position outside the quinoline moiety.

The examples in Figure 8 show how a PS can be embedded in different structural environments provided by distinct series of analogues and illustrate how relevant SAR information can be extracted from AC clusters.

CONCLUSIONS

We have explored AC formation by compounds containing PSs that have a long history in pharmaceutical research and have become focal points for generating target-class-directed compounds. Although prominent PSs are not class-specific, as also shown herein, compounds containing PSs are often enriched with specific biological activities. Some PSs are controversially viewed considering potential liabilities such as assay interference, but even these PSs are found in marketed drugs, indicating that compounds containing such PSs must be analyzed on a case-by-case basis. In our large-scale analysis, ACs containing PSs were frequently detected. Molecular properties of compounds forming PS-ACs were very similar to those forming other ACs. However, compounds from PS-ACs formed on average more AC relationships than others. Furthermore, substantial differences in AC formation between individual PSs were detected. Some PSs were present in much larger proportions of AC compounds than other PSs. Moreover, ACs containing selected PSs displayed unusually large improvements in ligand efficiency. We also emphasize that PSs are embedded in compounds in rather different ways and that their activity is structurally context-dependent. This was well accounted for by PS-AC network analysis, which separated clusters of ACs formed by different analogue series containing the same PS with activity against different targets. As we have shown, these clusters reveal target-specific SAR information for compound series containing a given PS. Thus, AC network analysis should aid in further exploring features of PSs that contribute to activity in different structural environments and determine target-based SARs. In addition to network analysis, ACs can also be further explored, for example, with the aid of complex X-ray structures, at least for structurally well-characterized targets, to examine individual interactions contributing to AC formation.

AUTHOR INFORMATION

Corresponding Author

Jürgen Bajorath − *Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität D-53115 Bonn, Germany;* orcid.org/0000-0002-0557-5714; Email: bajorath@bit.uni-bonn.de

Author

Huabin Hu − *Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität D-53115 Bonn, Germany*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.molpharmaceut.9b01236

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

H.H. is supported by the China Scholarship Council.

■ REFERENCES

(1) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J. Med. Chem.* **2012**, *55*, 2932−2942.

(2) Maggiora, G. M. On Outliers and Activity Cliffs–Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535−1535.

(3) Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57*, 18−28.

(4) Bajorath, J. Evolution of the Activity Cliff Concept for Structure-Activity Relationship Analysis and Drug Discovery. *Future Med. Chem.* **2014**, *6*, 1545−1549.

(5) Stumpfe, D.; Hu, H.; Bajorath, J. Evolving Concept of Activity Cliffs. *ACS Omega* **2019**, *4*, 14360−14368.

(6) Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. Matched Molecular Pairs as a Medicinal Chemistry Tool. *J. Med. Chem.* **2011**, *54*, 7739−7750.

(7) 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.

(8) 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.

(9) 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.

(10) Hu, H.; Stumpfe, D.; Bajorath, J. Rationalizing the Formation of Activity Cliffs in Different Compound Data Sets. *ACS Omega* **2018**, *3*, 7736−7744.

(11) Evans, B. E.; Rittle, K. E.; Bock, M. G.; DiPardo, R. M.; Freidinger, R. M.; Whitter, W. L.; Lundell, G. F.; Veber, D. F.; Anderson, P. S. Methods for Drug Discovery: Development of Potent, Selective, Orally Effective Cholecystokinin Antagonists. *J. Med. Chem.* **1988**, *31*, 2235−2246.

(12) Müller, G. Medicinal Chemistry of Target Family-Directed Masterkeys. *Drug Discovery Today* **2003**, *8*, 681−691.

(13) Horton, D. A.; Bourne, G. T.; Smythe, M. L. The Combinatorial Synthesis of Bicyclic Privileged Structure or Privileged Substructures. *Chem. Rev.* **2003**, *103*, 893−930.

(14) Welsch, M. E.; Snyder, S. A.; Stockwell, B. R. Privileged Scaffolds for Library Design and Drug Discovery. *Curr. Opin. Chem. Biol.* **2010**, *14*, 347−361.

(15) Schnur, D. M.; Hermsmeier, M. A.; Tebben, A. J. Are Target-Family-Privileged Substructures Truly Privileged? *J. Med. Chem.* **2006**, *49*, 2000−2009.

(16) Klabunde, T.; Hessler, G. Drug Design Strategies for Targeting G-Protein-Coupled Receptors. *ChemBioChem* **2002**, *3*, 928−944.

(17) Salaski, E. J.; Krishnamurthy, G.; Ding, W.-D.; Yu, K.; Insaf, S. S.; Eid, C.; Shim, J.; Levin, J. L.; Tabei, K.; Toral-Barza, L.; Zhang, W.-G.; McDonald, L. A.; Honores, E.; Hanna, C.; Yamashita, A.; Johnson, B.; Li, Z.; Laakso, L.; Powell, D.; Mansour, T. S. Pyranonaphthoquinone Lactones: A New Class of AKT Selective Kinase Inhibitors Alkylate a Regulatory Loop Cysteine. *J. Med. Chem.* **2009**, *52*, 2181−2184.

(18) Yan, A.; Wang, L.; Xu, S.; Xu, J. Aurora-A Kinase Inhibitor Scaffolds and Binding Modes. *Drug Discovery Today* **2011**, *16*, 260−269.

(19) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945−D954.

(20) *OEChem. TK*, version 1.7.7; OpenEye Scientific Software, Inc.: Santa Fe, NM, USA, 2012.

(21) Stumpfe, D.; Hu, H.; Bajorath, J. Introducing a New Category of Activity Cliffs with Chemical Modifications at Multiple Sites and Rationalizing Contributions of Individual Substitutions. *Bioorg. Med. Chem.* **2019**, *27*, 3605−3612.

(22) *RDKit: Cheminformatics and Machine Learning Software*, 2013; http://www.rdkit.org (accessed Apr 25, 2019).

(23) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. *Studies in Classification, Data Analysis, and Knowledge Organization* **2008**, 319−326.

(24) Abad-Zapatero, C.; Metz, J. T. Ligand Efficiency Indices as Guideposts for Drug Discovery. *Drug Discovery Today* **2005**, *10*, 464−469.

(25) Reynolds, C. H.; Bembenek, S. D.; Tounge, B. A. The Role of Molecular Size in Ligand Efficiency. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 4258−4261.

(26) Dahlin, J. L.; Nissink, J. W. M.; Strasser, J. M.; Francis, S.; Higgins, L.; Zhou, H.; Zhang, Z.; Walters, M. A. PAINS in the Assay: Chemical Mechanisms of Assay Interference and Promiscuous Enzymatic Inhibition Observed during a Sulfhydryl-Scavenging HTS. *J. Med. Chem.* **2015**, *58*, 2091−2113.

(27) Baell, J. B. Feeling Nature's PAINS: Natural Products, Natural Product Drugs, and Pan Assay Interference Compounds (PAINS). *J. Nat. Prod.* **2016**, *79*, 616−628.

(28) Jasial, S.; Hu, Y.; Bajorath, J. How Frequently Are Pan-Assay Interference Compounds Active? Large-Scale Analysis of Screening Data Reveals Diverse Activity Profiles, Low Global Hit Frequency, and Many Consistently Inactive Compounds. *J. Med. Chem.* **2017**, *60*, 3879−3886.

Summary

With the aid of a PS collection, a large-scale substructure search for PSs was conducted in the ChEMBL database. In this analysis, we only selected PSs that were contained in at least 100 distinct bioactive compounds, yielding 24 PSs for further analysis. These 24 PSs were found in around 42% of all high-confidence bioactive compounds, indicating their high utilization as templates for compound design. We also reevaluated the hypothesis that these PSs were not truly target family-privileged substructures. Instead, most of them were found in distinct target families. Three different types of ACs, i.e., MMP-clifs, isomer cliffs and dual-site ACs, were systematically generated and searched for the presence of PSs. ACs with and without PSs showed similar distributions with respect to total AC numbers. Moreover, no significant differences in molecular properties (e.g., ligand efficiency, logP, PSA and sp^3 carbon) between ACs with and without PSs were observed. When different PSs were compared, some PSs showed a high propensity to form ACs. For PSs detected in different activity classes, PS-based AC networks were constructed which showed that the same PS embedded in different structural contexts displayed distinct activities, as reflected by disjoint AC clusters.

ACs have been generally underutilized in medicinal chemistry. In the next chapter, a unified strategy for extracting different types of ACs is introduced. These ACs are made available to the public for subsequent analysis and extended AC data structure for rationalizing dual-site ACs is suggested.

Chapter 5

Increasing the Public Activity Cliff Knowledge Base with New Categories of Activity Cliffs

Introduction

In the previous chapter, different types of ACs were computationally generated. The comparison of ACs with and without PSs indicated that they were difficult to differentiate based on molecular properties. However, the drug-likeness and promiscuous manner of these PSs will continue to be highly attractive for bioactive compound design.

In this chapter, dual-site cliffs, isomer cliffs, and ACs containing PSs (PS-ACs) are described and made available to the public. Additionally, for dual-site ACs, a practical approach for extracting SAR information is suggested, thus making them more enticing for medicinal chemistry.

Data Note

For reprint orders, please contact: reprints@future-science.com

Future Science OA

Increasing the public activity cliff knowledge base with new categories of activity cliffs

Huabin Hu[1] & Jürgen Bajorath*[,1]

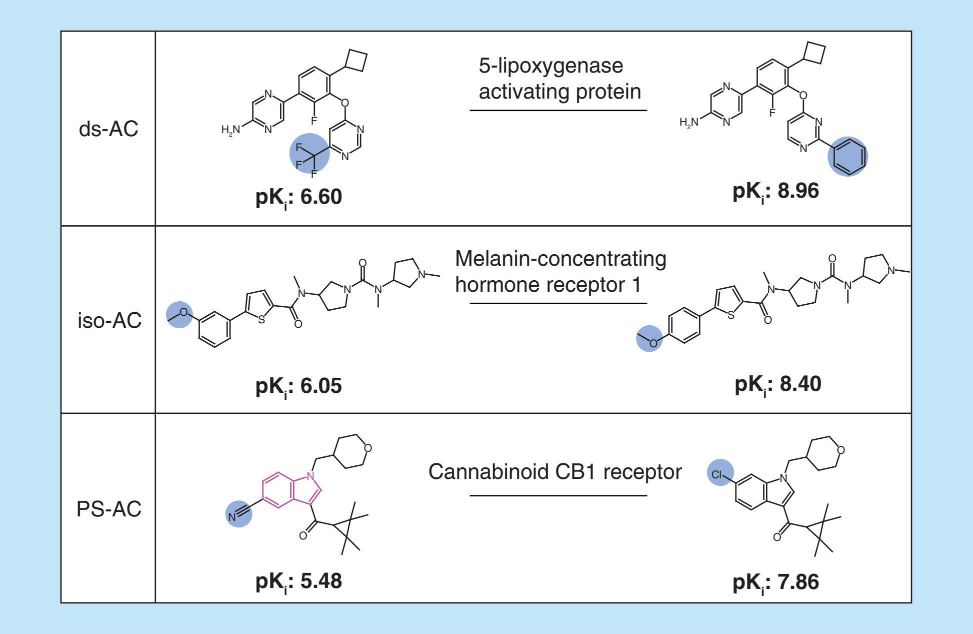
[1] Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology & Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Endenicher Allee 19c, Bonn D-53113, Germany

*Author for correspondence: Tel.: +49 228 7369 100; Fax: +49 228 7369 101; bajorath@bit.uni-bonn.de

Aim: Extending the public knowledge base of activity cliffs (ACs) with new categories of ACs having special structural characteristics. **Methodology:** Dual-site ACs, isomer ACs and ACs with privileged substructures are described and their systematic identification is detailed. **Exemplary results & data:** More than 7400 new ACs belonging to different categories with activity against more than 200 targets were identified and are made publicly available. **Limitations & next steps:** For dual-site ACs, limited numbers of isomers are available as structural analogs for rationalizing contributions to AC formation. The search for such analogs will continue. In addition, the target distribution of ACs containing privileged substructures will be further analyzed.

Lay abstract: Activity cliffs (ACs) are formed by small molecules that have very similar structures, are active against the same biological target, but have a large difference in potency against their target. Accordingly, ACs are of interest in medicinal chemistry because they reveal small structural changes that greatly influence the potency of active compounds. This information can be used for compound optimization. Computational methods are applied to search for ACs in large compound databases. Here, we further extend the public AC knowledge base with new categories of ACs having special structural characteristics.

Graphical abstract: Shown are an exemplary dual-site activity cliff (AC) (top), isomer AC (middle) and an AC containing a privileged substructure (bottom). Structural differences between AC compounds are highlighted in blue and the privileged substructure is colored pink. Compound targets and potency (pK_i) values are reported.



First draft submitted: 9 February 2020; Accepted for publication: 4 March 2020; Published online: 15 April 2020

Keywords: active compounds • activity cliffs • computational analysis • medicinal chemistry • open access data • privileged substructures • structural isomers • substitution sites

Activity cliffs (ACs) are defined as pairs or groups of structurally similar (analogous) compounds that are active against the same target but have a large difference in potency [1–4]. ACs have also been studied on the basis of compounds, which are highly potent against a given target and structural analogs that are confirmed inactive against this target [4]. Furthermore, ACs have been investigated from a variety of research perspectives including the consideration of different AC concepts, different types of data analysis and AC predictions [1–9]. In medicinal chemistry, ACs are of particular interest because they capture small chemical modifications of active compounds that substantially contribute to, or determine, structure–activity relationships (SARs) [2,3].

For formally defining ACs, molecular similarity and potency difference criteria must be specified [2–4]. Similarity can be calculated on the basis of chemical descriptors and numerical similarity metrics (descriptor-based/numerical similarity) or on the basis of substructure relationships (substructure-based similarity) [3,4]. Substructure-based similarity measures include shared scaffolds, the formation of matched molecular pairs (MMPs) or membership in the same analog series (AS) [4,10,11]. Compounds forming MMP-based ACs are confined to chemical changes at a single substitution site [10,12], whereas AS-based ACs may contain single or multiple substitution sites [11,13].

Furthermore, for defining ACs, constant potency difference thresholds can be applied across different compound activity classes (e.g., at least 100-fold potency difference) [2,3]. Alternatively, activity class-dependent potency difference thresholds can be determined on the basis of statistically significant potency differences, with respect to intra-class potency value distributions [14,15]. In either case, the use of assay-independent equilibrium constants (K_i values) as potency measurements is generally preferred over assay-dependent measurements such as IC_{50} values. The use of K_i values makes it possible to compare ACs for a given target and across different targets in a meaningful way.

Considering the evolution of the AC concept in medicinal chemistry [4], we have defined three generations of ACs [4,16], depending on the structural similarity and potency difference criteria that are applied:

First generation ACs

Similarity criterion: numerical or substructure-based similarity.

Potency difference criterion: constant potency difference threshold across all activity classes.

Second generation ACs

Similarity criterion: MMP formation (analog pairs with single substitution site).

Potency difference criterion: variable activity class-dependent potency difference thresholds.

Third generation ACs

Similarity criterion: structural analogs originating from the same AS (with single or multiple substitution sites).

Potency difference criterion: variable activity class-dependent potency difference thresholds.

Previously, we have generated a large collection of second generation ACs [15,17] and made it publicly available as an open access deposition [17,18].

For nearly 100 different activity classes, each representing a unique target protein, more than 20,000 activity class-dependent ACs were identified, also taking structural analogs of potent compounds into account that were confirmed inactive against the same target [15,17]. Compound activity data were extracted from medicinal chemistry sources (ChEMBL database) [19] and high-throughput screens (PubChem Bioassays) [20].

Herein, we further increase the public AC knowledge base through addition of three recently introduced categories of ACs including dual-site ACs (ds-ACs) [13], isomer ACs (iso-ACs) [21] and ACs containing ‘privileged substructures’ (PS-ACs) [22]. These AC categories are detailed in the methodology section and an in-depth analysis of each category is reported in its original publication.

The PS concept was originally introduced in medicinal chemistry by Evans *et al.* [23] and has become increasingly popular over time [24,25]. PSs are frequently found in compounds with preferential activity against specific target families. They usually are not selective for a particular target but display a tendency of preferential binding to

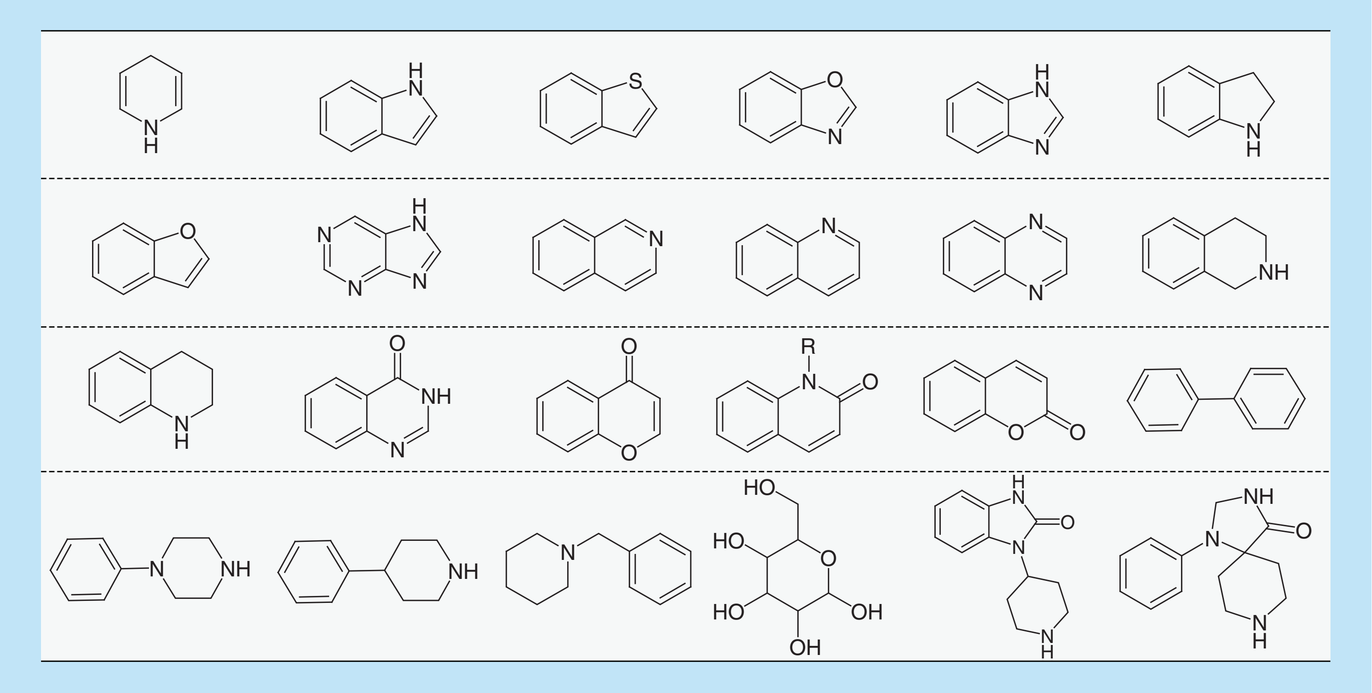


Figure 1. Privileged substructures. Shown are the structures of 24 privileged substructures found in at least 100 unique bioactive compounds.

an individual target family over others. Accordingly, PSs are often considered as family-directed core structures in medicinal chemistry [23–25]. Studying PS-ACs is attractive because these ACs reveal different levels of SAR information associated with individual PSs, as described in detail [22].

In the following, we report a systematic search for ds-ACs, iso-ACs and PS-ACs, resulting in a new collection of ACs that further extends our knowledge base of ACs and is made available as an open access deposition. Importantly, ds-ACs, iso-ACs and PS-ACs were originally introduced in independent studies. Herein, we report a new unified search strategy that has made it possible to identify these ACs in bioactive compounds in concert applying consistent criteria, determine the overlap between different AC categories and study ACs belonging to these categories. This strategy is related to, yet distinct from the one applied in the original assessment of PS-ACs, which were most recently introduced [22], and has yielded the first public collection of PS-ACs. All new ACs identified in our systematic analysis are made freely available as a part of this study, providing a wealth of examples for follow-up investigations.

Methodology

Compound activity data

Bioactive compounds were extracted from the latest version of the ChEMBL [19] database (release 25). For selection of high-confidence activity data, rigorous criteria were applied. Only compounds forming direct interactions with human targets (i.e., assay relationship type ‘D’) at the highest assay confidence level (i.e., assay confidence score 9) were selected. Furthermore, only equilibrium constants (i.e., K_i values) with specified numerical values (‘=’ relationship) were accepted as potency measurements for given targets.

Privileged substructures

PSs were defined according to Welsch *et al.* [25]. A systematic search was carried out for PSs that were contained in 100 or more unique ChEMBL compounds. Figure 1 shows 24 PSs that were identified and further considered for AC analysis.

Compound fragmentation

To ensure consistent generation of ACs belonging to different categories, a recently introduced compound frag-

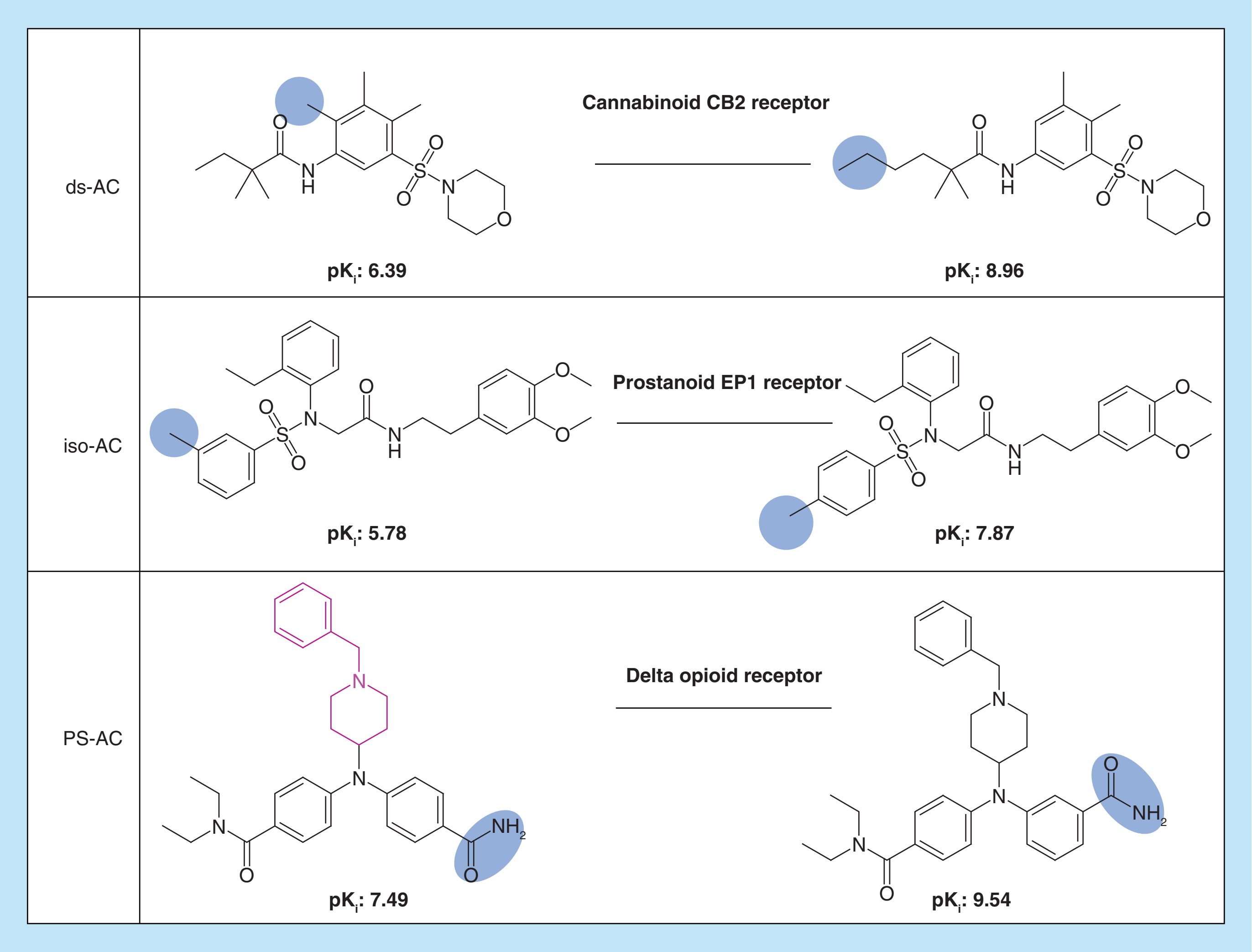


Figure 2. Exemplary activity cliffs belonging to different categories. From the top to the bottom, an exemplary ds-AC, iso-AC and PS-AC are shown. For each AC, the target and compound potency (pK$_i$) values are reported. Structural modifications and the PS are colored blue and pink, respectively.

AC: Activity cliff; ds-AC: Dual-site activity cliff; iso-AC: Isomer activity cliff; PS: Privileged substructure.

mentation scheme was applied [22]. Using a decomposition algorithm for MMP generation [10], exocyclic single bonds in compounds were systematically fragmented, yielding two fragments per step. During the fragmentation process, the following size restrictions were applied to obtain a core and substituent fragment [12]. The number of nonhydrogen atoms of the core fragment was required to be at least twice as large as the number of nonhydrogen atoms comprising the substituent fragment. In addition, the size of the substituent was restricted to at most 13 nonhydrogen atoms. Furthermore, the substituent in the core fragment was replaced by a hydrogen atom (R → H). The calculations were carried out using in-house scripts with the aid of the OpenEye chemistry toolkit (NM, USA) (version 1.7.7) [26].

Analog pairs & activity cliffs

Following fragmentation, compounds having the same activity and sharing the same core were organized into individual sets of analogs. Then, analog pairs (APs) differing at two substitution sites were systematically enumerated and categorized as follows:

- Structural isomers: the same substituent occurred at two different sites (core positions).

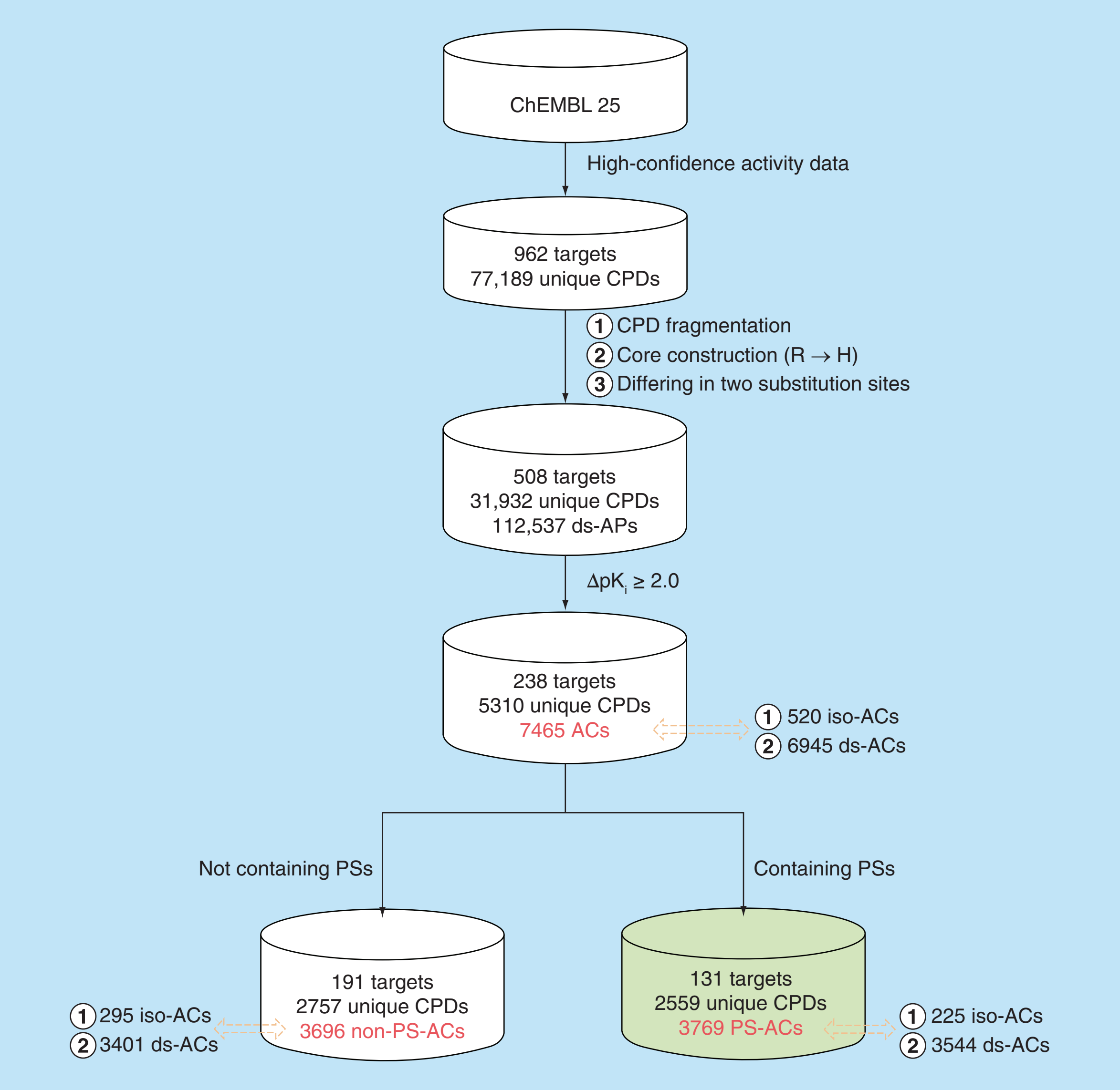


Figure 3. Unified search strategy for the identification of different activity cliffs. The identification of ds-ACs, iso-ACs and PS-ACs is summarized. Numbers of compounds, targets and ACs are given at each stage.
AC: Activity cliff; CPD: Compound; ds-AC: Dual-site activity cliff; ds-AP: Dual-site analog pair; iso-AC: Isomer activity cliff; PS: Privileged substructure; PS-AC: Privileged substructure-containing activity cliff.

- Dual-site analogs: two different substituents occurred at different sites. The size difference between these exchanged substituents was restricted to at most eight non-hydrogen atoms.

For each AP, it was determined whether the participating compounds had an at least 100-fold difference in potency, which qualified the pair as an iso-AC or ds-AC. We note that iso-ACs are confined to structural isomers and hence distinct from chirality or chiral cliffs [4,9] where cliff compounds are distinguished by different chirality at a given stereocenter. Furthermore, for each AC, it was determined if it contained a PS. ACs with PSs were also classified as PS-ACs. Figure 2 shows exemplary ACs belonging to different categories. By definition, iso-ACs represent a special case of ds-ACs.

Detection of isomers of dual-site activity cliff compounds

For ds-AC compounds, a further systematic search for structural isomers (from the same activity class) with the same substituent at the other substitution site was carried out. If such isomers were identified, it was possible

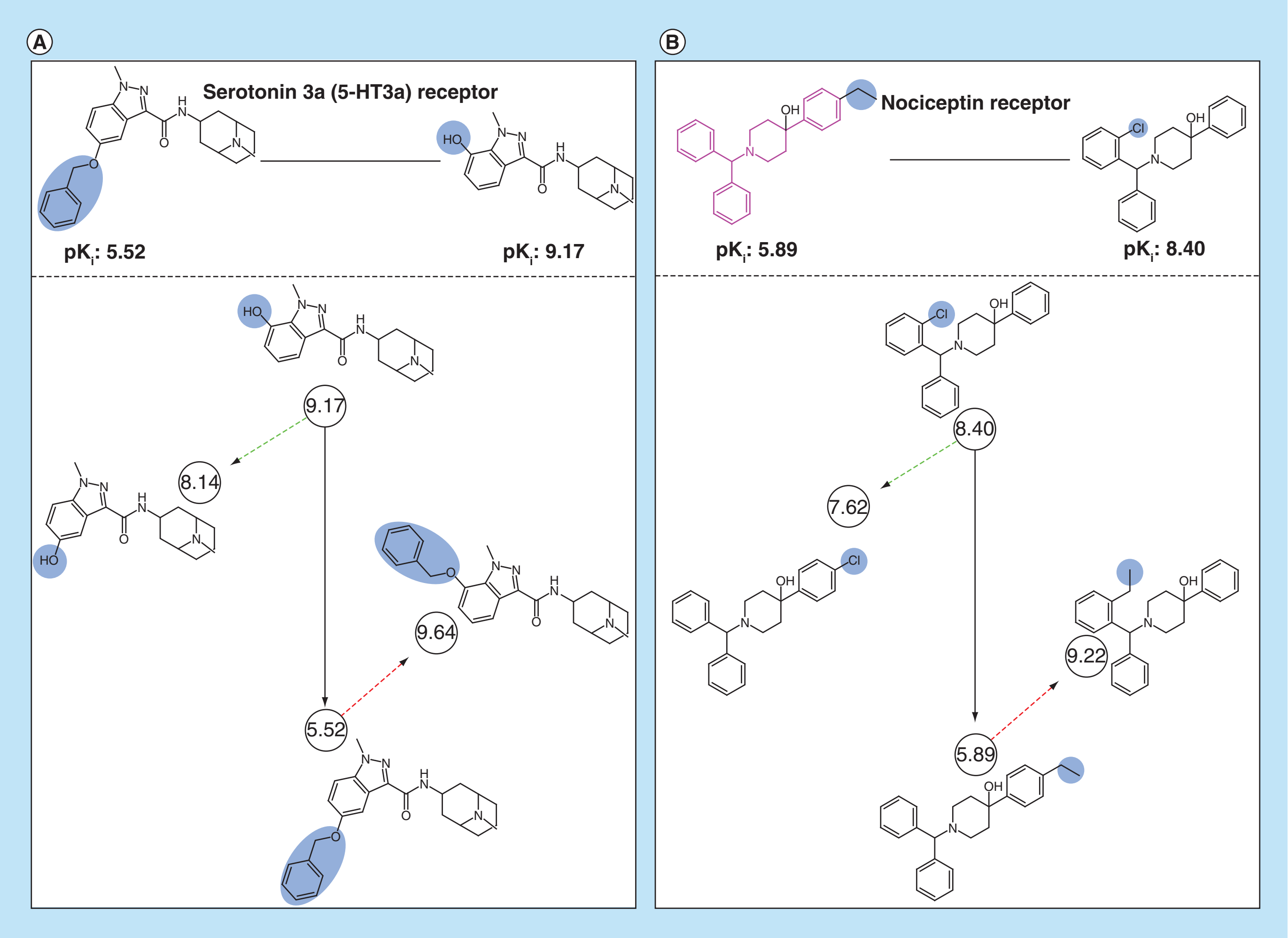


Figure 4. Extended data structure for dual-site activity cliffs. Shown are exemplary ds-ACs and corresponding data structures **(A)** without and **(B)** with a PS. Highly and weakly potent ds-AC compounds are connected by solid black arrows. Structural isomers of highly and weakly potent AC compounds are connected to corresponding ds-AC compounds using dashed green and red arrows. Structural modifications and the PS are colored blue and pink, respectively. For each example, the target name and compound potency (pK_i) values (in circles) are reported.
AC: Activity cliff; ds-AC: Dual-site activity cliff; PS: Privileged substructure.

to generate an extended data structure for a ds-AC, revealing the contributions of substituent positions to AC formation, as further discussed below.

Exemplary results & data

Unified search strategy for activity cliffs belonging to different categories

Originally, ds-ACs, iso-ACs and PS-ACs were separately studied. Here, we have implemented a unified search strategy to identify ACs belonging to these categories in parallel and determine their overlap. The search strategy is summarized in Figure 3. After compound fragmentation, a total of 112,537 qualifying APs were identified that yielded a total of 7465 ACs, which were assigned to different categories, as further detailed below.

Extended dual-site activity cliff data structure

For SAR exploration, ds-ACs can be extended to generate a data structure comprising four analogs, as illustrated in Figure 4. This data structure makes it possible to examine the contributions of substituent positions to ds-AC formation and is thus rich in SAR information [13]. Its generation requires the identification of isomers of ds-AC compounds with inversely repositioned substituents, as shown in Figure 4. A systematic search for such isomers revealed that 396 ds-ACs could be fully extended with two qualifying isomers. In addition, for 2320 other ds-ACs, only one isomer was identified. Among analogs from different series, structural isomers are statistically under-represented when compared with analogs carrying different substituents. A possible reason might be that medicinal chemists, from an SAR perspective, typically prefer introducing different substituents at a given site, rather than synthesizing analogs with a 'moving' substituent (structural isomers). Regardless, the extended data structure based upon ds-ACs offers additional opportunities for SAR analysis and illustrates the utility of this AC category.

Data

Our systematic search identified a total of 3696 ACs without PSs that were formed by 2757 unique compounds with activity against 191 targets. These ACs included 3401 ds- and 295 iso-ACs. Thus, only a limited number of iso-ACs were available. In addition, the search identified 3769 PS-ACs formed by 2559 unique compounds with activity against 131 targets. These PS-ACs included 3544 ds- and 225 iso-ACs. ACs with and without PSs shared 84 targets.

Our analysis revealed that approximately half of the newly identified ACs contained one of 24 predefined PSs that were detected in at least 100 unique bioactive compounds. The high frequency with which a predefined set of PSs occurred in ds- and iso-ACs, thus combining different AC categories, indicated that PSs yielded SAR-informative compounds with potential for further optimization. Hence, on the basis of AC analysis, these PSs deserve further consideration in medicinal chemistry. The PS-ACs provided as a part of our study should aid in further exploring these PSs.

Data deposition

All ACs identified herein are provided in a text file. For each AC, category membership(s) are specified. For AC compounds, the ChEMBL ID, canonical SMILES representation and potency value are reported. For PS-ACs identified herein (forming a subset of iso-ACs and ds-ACs), the SMILES string of the PS is also provided. The data are made freely available as a deposition on the ZENODO open access platform [27].

Limitations & next steps

The extended ds-AC data structure enables the analysis of substitution site-specific contributions to AC formation. However, among structural analogs, structural isomers are under-represented and only limited numbers of isomers are currently available for ds-AC analysis. This is essentially the only data-dependent limitation associated with exploring the new AC categories introduced here. Hence, the search for isomers as structural analogs for ds-AC analysis will continue. Furthermore, the large number of PS-ACs we identified makes it possible to investigate the target distribution and SAR information content of PS-containing compounds and their analogs in greater detail. For this purpose, PS-ACs provide immediate focal points.

Executive summary

- The activity cliffs (AC) concept is rationalized.
- Different generations of ACs are defined.

Methodology

- Procedures for AC identification are detailed.
- Recently introduced AC categories are described.
- Search routines are implemented.

Exemplary results & data

- A unified search strategy for identifying different ACs is detailed.
- Search results are summarized.
- An extended data structure based upon dual-site ACs is introduced.
- A collection of ACs is generated.
- Details of its open access deposition are provided.

Limitations & next steps

- Limited availability of isomers of dual-site AC compounds is discussed.
- Further analysis of privileged substructure-containing ACs is proposed.

Author contributions

J Bajorath and H Hu conceived the study; H Hu carried out the analysis; H Hu and J Bajorath analyzed the results, J Bajorath and H Hu prepared the manuscript.

Acknowledgments

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

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

Open access

References

Papers of special note have been highlighted as: • of interest; •• of considerable interest

1. Maggiora GM. On outliers and activity cliffs – why QSAR often disappoints. *J. Chem. Inf. Model* 46(4), 1535–1535 (2006).

•• **First explicit discussion of activity cliffs (ACs) in the chemical literature.**

2. Stumpfe D, Bajorath J. Exploring activity cliffs in medicinal chemistry. *J. Med. Chem.* 55(7), 2932–2942 (2012).

• **First review of the AC concept in medicinal chemistry.**

3. Stumpfe D, Hu Y, Dimova D, Bajorath J. Recent progress in understanding activity cliffs and their utility in medicinal chemistry. *J. Med. Chem.* 57(1), 18–28 (2014).

4. Stumpfe D, Hu H, Bajorath J. Evolving concept of activity cliffs. *ACS Omega* 4(11), 14360–14368 (2019).

• **Most recent review of AC research and extension of the AC concept.**

5. Cruz-Monteagudo M, Medina-Franco JL, Pérez-Castillo Y, Nicolotti O, Cordeiro MNDS, Borges F. Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? *Drug Discov. Today* 19(8), 1069–1080 (2014).

6. Medina-Franco JL. Activity cliffs: facts or artifacts? *Chem. Biol. Drug. Des.* 81(5), 553–556 (2013).

7. Medina-Franco JL. Scanning structure–activity relationships with structure–activity similarity and related maps: from consensus activity cliffs to selectivity switches. *J. Chem. Inf. Model.* 52(10), 2485–2493 (2012).

8. Pérez-Benito L, Casajuana-Martin N, Jiménez-Rosés M, van Vlijmen H, Tresadern G. Predicting activity cliffs with free-energy perturbation. *J. Chem. Theory Comput.* 15(3), 1884–1895 (2019).

9. Schneider N, Lewis RA, Fechner N, Ertl P. Chiral cliffs: investigating the influence of chirality on binding affinity. *ChemMedChem* 13(13), 1315–1324 (2018).

10. Hussain J, Rea C. Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *J. Chem. Inf. Model* 50(3), 339–348 (2010).

11. Naveja JJ, Vogt M, Stumpfe D, Medina-Franco JL, Bajorath J. Systematic extraction of analogue series from large compound collections using a new computational compound–core relationship method. *ACS Omega* 4(1), 1027–1032 (2019).

12. 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* 52(5), 1138–1145 (2012).

13. Stumpfe D, Hu H, Bajorath J. Introducing a new category of activity cliffs with chemical modifications at multiple sites and rationalizing contributions of individual substitutions. *Bioorg. Med. Chem.* 27(16), 3605–3612 (2019).

14. Hu H, Stumpfe D, Bajorath J. Rationalizing the formation of activity cliffs in different compound data sets. *ACS Omega* 3(7), 7736–7744 (2018).

15. Hu H, Stumpfe D, Bajorath J. Second-generation activity cliffs identified on the basis of target set-dependent potency difference criteria. *Future Med. Chem.* 11(5), 379–394 (2019).

•• **Introduction of ACs with variable potency difference thresholds.**

16. Stumpfe D, Hu H, Bajorath J. Computational method for the identification of third generation activity cliffs. *MethodsX* 7, 100793 (2020).

17. Hu H, Stumpfe D, Bajorath J. Systematic identification of target set-dependent activity cliffs. *Future Sci. OA* 5(2), FSO363 (2019).

18. Hu H, Stumpfe D, Bajorath J. Target set-dependent activity cliffs (2018). https://doi.org/10.5281/zenodo.1436584

19. Gaulton A, Bellis LJ, Bento AP *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40(D1), D1100–D1107 (2012).

20. Wang Y, Bryant SH, Cheng T *et al.* PubChem BioAssay: 2017 update. *Nucleic Acids Res.* 45(D1), D955–D963 (2017).

21. Hu H, Bajorath J. Introducing a new category of activity cliffs combining different compound similarity criteria. *RSC Med. Chem.* 11(1), 132–141 (2020).

22. Hu H, Bajorath J. Systematic exploration of activity cliffs containing privileged substructures. *Mol. Pharmaceutics* 17(3), 979–989 (2020).

• **Systematic identification of privileged substructure-ACs.**

23. Evans BE, Rittle KE, Bock MG *et al.* Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. *J. Med. Chem.* 31(12), 2235–2246 (1988).

•• **Introduction of the privileged substructure concept.**

24. Müller G. Medicinal chemistry of target family-directed masterkeys. *Drug Discov. Today* 8(15), 681–691 (2003).

25. Welsch ME, Snyder SA, Stockwell BR. Privileged scaffolds for library design and drug discovery. *Current Opin. Chem. Biol.* 14(3), 347–361 (2010).

• **Comprehensive compendium of known privileged substructures.**

26. OEChem TK. Version 1.7.7. OpenEye Scientific Software, Inc, NM, USA (2012). www.eyesopen.com

27. Hu H, Bajorath J. New categories of activity cliffs (2020). https://doi.org/10.5281/zenodo.3660200

Summary

Herein, through a unified strategy, 6945 dual-site cliffs and 520 isomer cliffs were obtained, which were formed by a total of 5310 compounds with activity against more than 200 targets. 3769 ACs ($\sim$ 50.4%) contained at least one PS implying that PSs were frequently used for SAR exploration. For these dual-site cliffs, a systematic structural isomer search for highly and weakly potent AC compounds was performed. Only 396 dual-site cliffs could be fully extended with two isomers with inversely repositioned substituents. For 2320 ACs, only one structural isomer of either the highly or weakly potent AC compound was identified. For the remainder, no structural isomers were detectable. These observations implied that structural analogs forming isomer relationships with AC compounds were generally underrepresented. The isomers in extended dual-site ACs helped to rationalize such cliffs and demonstrated the utility of this AC category. Associated ACs were made available to the public in an open-access deposition.

Since ACs encode unexpected biological responses, they might be involved in critical protein-ligand interactions. If these could be identified, they would be highly informative for structure-based drug design. In the next chapter, ACs with minimal chemical modifications are systematically extracted and rationalized with the aid of X-ray structures.

Chapter 6

Activity Cliffs Produced by Single-Atom Modification of Active Compounds: Systematic Identification and Rationalization Based on X-Ray Structures

Introduction

Molecular recognition largely depends on various favorable interactions such as hydrogen bond formation or covalent bonding. In drug discovery, it is highly appreciated when minor structural modifications are accompanied with increase in activities, yielding promising candidates. Such effects indicate specific SAR determinants, which are of particular interest in structure-based drug design.

In this chapter, we introduced computational methods to systematically extract ACs that captured minimal structural modifications including heteroatom replacement and positional difference. Since the total number of heavy atoms between AC compounds remains the same, the formation of this AC category indicates a ligand efficiency improvement that originates purely from potency effects. For these newly identified ACs, a search for X-ray complexes was performed to rationalize the AC formation at the atomic level.

European Journal of Medicinal Chemistry 207 (2020) 112846

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

journal homepage: http://www.elsevier.com/locate/ejmech

Activity cliffs produced by single-atom modification of active compounds: Systematic identification and rationalization based on X-ray structures

Huabin Hu, Jürgen Bajorath*

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

ARTICLE INFO

Article history:
Received 13 August 2020
Received in revised form
10 September 2020
Accepted 11 September 2020
Available online 16 September 2020

Keywords:
Active compounds
Activity cliffs
Single-atom modification
Lead optimization
Activity data analysis
X-ray structures

ABSTRACT

In medicinal chemistry, activity cliffs (ACs) are considered as sources of critical structure-activity relationship (SAR) information. ACs are capable of revealing such SAR information because they are formed by pairs or groups of structural analogs that are distinguished by small chemical modifications leading to large variations in compound potency. Such modifications can reveal critically important substitution sites in analog series. Small AC-encoded chemical changes enable the identification of SAR determinants. In this work, we have searched medicinal chemistry data for most "subtle" ACs in which participating compounds are only distinguished by single-atom modifications. These ACs can be directly associated with lead optimization strategies such as positional atom scanning (atom "walks") or heteroatom replacements in ring structures. More than 1500 of these ACs with activity against a variety of targets were identified. To further explore newly identified ACs, we searched for X-ray structures of ligand-target complexes containing participating AC compounds. For a subset of subtle ACs, X-ray structures of complexes made it possible to examine effects of single-atom changes in light of well-defined ligand-target interactions. Since ACs capturing minimal chemical changes are of particular interest for lead optimization and drug design, we make all newly identified ACs and associated structural information freely available as an open access deposition.

1. Introduction

Compound optimization depends on the exploration of structure-activity relationship (SAR) information that is typically collected while generating series of analogs [1]. Alternatively, going beyond individual analog series, SAR information can be obtained through systematic analysis of compound activity data [2]. To capture critical SAR information provided by analog series or large compound data sets, the activity cliff (AC) concept was introduced [3,4]. ACs were originally defined as pairs of structurally similar active compounds with large potency differences [3]. As such, ACs can be understood as the pinnacle of SAR discontinuity in compound series. Accordingly, their occurrence is typically desirable during the early stages of compound optimization when compound potency must be improved and detection of potency gradients is desirable [5]. By contrast, during late stages of lead optimization when high potency levels must be retained while balancing multiple compound properties, encountering steep SARs, as indicated by ACs, is less desirable [5]. Regardless, ACs have high SAR information content and often reveal SAR determinants, the notion of which aids in guiding compound optimization efforts.

Over time, the general pair-based definition of ACs has evolved and molecular similarity and potency difference criteria underlying AC assessment have been further refined [3,6]. In addition, alternative AC representations have been considered ranging from molecular graph-based feature sets to bioactive compound conformations [6]. As a source of knowledge for practical applications in medicinal chemistry, ACs are typically most useful if they are represented by structural analogs with well-defined substitution sites [6]. In addition, simple compound modifications increase the chemical interpretability of ACs. The smaller AC-inducing substitutions are, the more likely they are to identify sites in analogs that determine SARs. Formation of ACs has also been observed for stereoisomers and structural isomers [7]. However, in these cases,

* Corresponding author.
E-mail address: bajorath@bit.uni-bonn.de (J. Bajorath).

https://doi.org/10.1016/j.ejmech.2020.112846
0223-5234/© 2020 Elsevier Masson SAS. All rights reserved.

AC-inducing effects are often difficult to understand without additional data.

We have been interested in exploring ACs that encode minimal chemical modifications: single-atom modifications. Therefore, we have conducted a systematic search for ACs with single-atom modifications. Such "subtle" ACs are of particular interest for medicinal chemistry, for two reasons. First, since these ACs contain minimal chemical modifications there is a high probability that they reveal SAR determinants. Second, these AC can be directly associated with advanced lead optimization strategies including positional atom scanning, also referred to as atom "walks", or heteroatom replacements in ring structures [8,9]. In addition, such ACs are also of particular interest as test cases for computational chemistry, especially for calibrating or benchmarking (free) energy methods or scoring functions. Therefore, we have systematically identified such ACs across current compound activity classes on the basis of curated high-confidence activity data. ACs capturing minimal chemical modifications are also of interest for structure-based compound design because they are likely to single out individual interactions that are important for ligand binding. Hence, we have searched X-ray structures of ligand-target complexes for AC targets and compounds and studied key interactions in detail, as also reported herein. As a part of our study, the newly identified ACs are made freely available for medicinal and computational chemistry applications.

2. Experimental

2.1. Compounds and activity data

From ChEMBL [10] version 26, all bioactive compounds meeting the following criteria were extracted. Only compounds forming direct interactions (assay relationship type: "D") with human targets at the highest confidence level (assay confidence score: 9) were selected. As potency measurements, only numerically specified equilibrium constants (K_i values) or IC_{50} values were considered and separately organized into two data sets, i.e., a K_i- and IC_{50}-based set. Approximate potency measurements with ">", "<" or "~" relationships were removed. If a compound had multiple potency values, the geometric mean of all values was calculated as the final potency annotation, provided all values fell within the same order of magnitude; otherwise, the compound was disregarded.

On the basis of these stringent selection criteria, a total of 85,598 unique bioactive compounds forming 144,356 interactions with 993 targets were obtained for the K_i–based data set. In addition, the IC_{50}–based set consisted of a total of 225,498 unique compounds with activity against 1841 targets, forming a total of 317,474 interactions.

Because K_i and IC_{50} measurements cannot be directly compared, analog pairs (APs) with single-atom modifications and corresponding ACs were separately identified for these two data sets, as described in the following.

2.2. Analog pairs and activity cliffs with single-atom modifications

On the basis of these high-confidence data sets, analog pairs (APs) with single-atom modifications were systematically identified with the aid of RDKit [11] and organized into corresponding AP sets.

Two compounds were paired if they:

(i) contained a single-atom replacement. Four different types of atom replacements were considered: N to C (N–C), O–C, N–O, and S–O. Such analogs formed an *atom-replacement* AP.
(ii) were distinguished only by the position of a single heteroatom, forming an *atom-walk* AP.

The IC_{50} and K_i sets yielded a total of 36,526 qualifying APs with activity against 1046 targets and 17,526 qualifying APs with activity against 489 targets, respectively. These sets of APs with single-atom modifications provided the basis for the identification of corresponding ACs.

An AP was classified as an AC if the two participating analogs displayed an at least 100-fold difference in potency. A potency difference of at least two orders of magnitude has frequently been generally applied as an AC criterion across different compound activity classes (also termed target sets) [3,6].

2.3. Activity cliff networks

To study the formation of newly identified ACs in context, AC networks were generated in which nodes represented participating AC compounds and edges pairwise AC relationships [12]. Partly overlapping ACs formed by groups of analogs with large potency variations are referred to as coordinated ACs [6,12]. In AC networks, the presence of coordinated ACs gives rise to the formation of clusters, which are rich in SAR information [12] and can be individually selected and studied. Network representations were drawn with Cytoscape [13].

2.4. Target classification and X-ray structures with activity cliff compounds

AC targets were assigned to different target classes and groups according to the ChEMBL l1 and l2 target/protein family classification scheme, respectively [10].

The RCSB Protein Data Bank (PDB) [14] was searched for X-ray structures of AC targets in complex with AC compounds identified herein. For AC targets, ChEMBL identifiers (IDs) were mapped to UniProt [15] and UniProt IDs were then used to search for PDB entries with the aid of KNIME protocols [16]. X-ray structures of AC targets were searched for co-crystallized AC compounds.

Ligand-target interactions in crystal structures with AC compounds were determined and visualized using protein structure analysis functions of the Molecular Operating Environment (MOE) [17,18].

Table 1
Activity cliff statistics.

Type	IC_{50} data set		K_i data set	
	Atom-replacement	Atom-walk	Atom-replacement	Atom-walk
N–C	396 (15,136)	256 (8182)	176 (6792)	113 (3782)
O–C	119 (4990)	25 (956)	61 (2435)	8 (467)
N–O	162 (4388)	3 (96)	110 (2170)	3 (78)
S–O	45 (2754)	0 (24)	37 (1779)	0 (23)
Total number	722 (27,268)	284 (9258)	384 (13,176)	124 (4350)

For each data set, the number of atom-replacement and atom-walk ACs and corresponding number of APs (in parentheses) are reported.

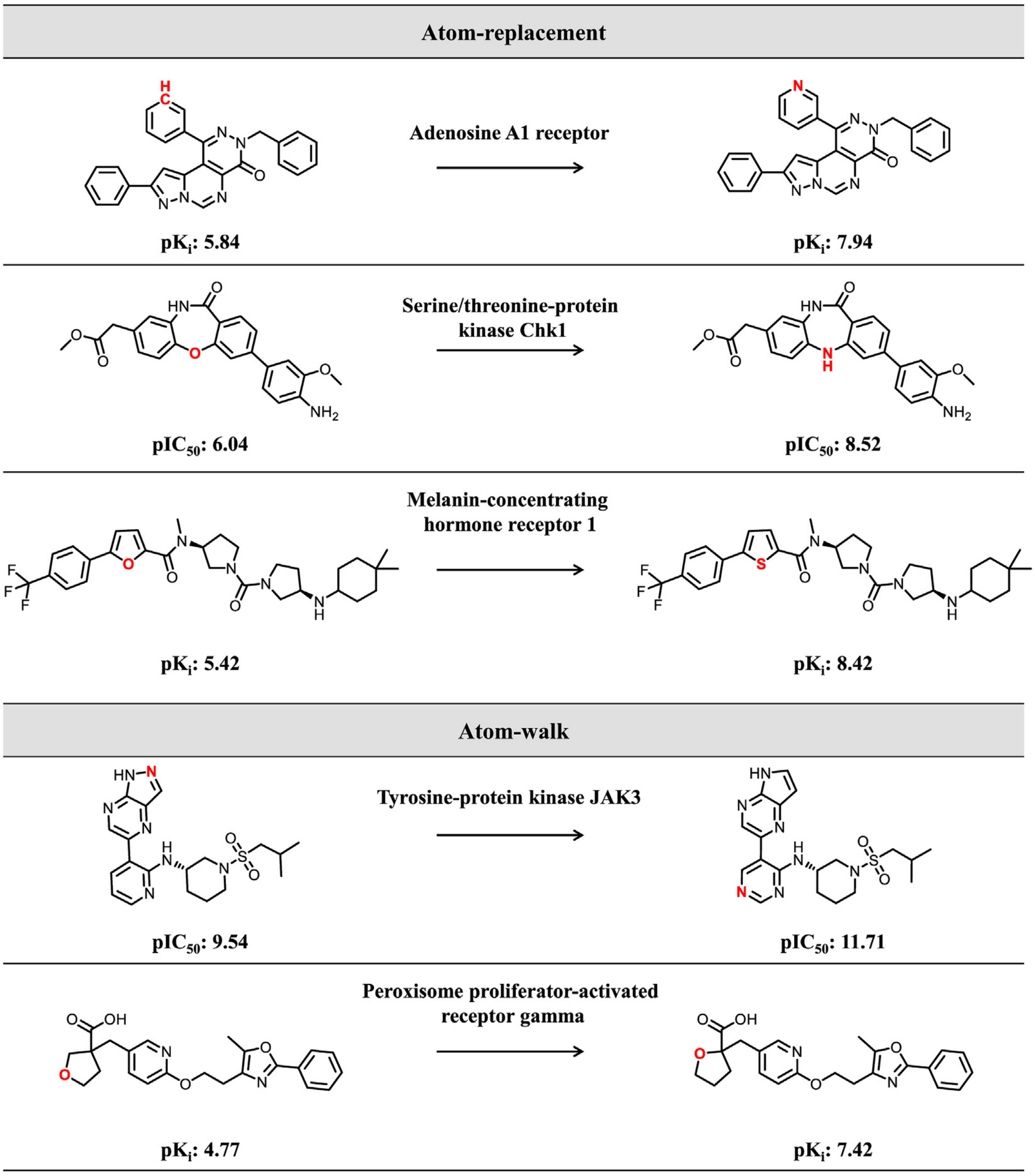


Fig. 1. Activity cliff categories. Shown are exemplary atom-replacement (top) and atom-walk ACs (bottom). In each case, the chemical modification is highlighted in red. For each AC, negative logarithmic potency (pK_i or pIC_{50}) values and the target protein are reported. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

2.5. Computational methods summary

The approach applied herein is summarized as follows: Initially, large volumes of compounds and activity data were curated to focus the analysis on high-confidence data, which is essential for AC analysis. Then, pairs of structural analogs with activity against the same target and well-defined single-atom modifications were systematically identified. From these APs, ACs were selected in which the two participating analogs had an at least 100-fold difference in potency. The formation of these ACs was then visualized and studied in detail in network representations in which nodes represented AC compounds and edges pairwise AC relationships. Finally, a search was carried out for X-ray structures of complexes between AC target proteins and AC compounds. Ligand-target interactions in these crystallographic structures were analyzed to aid in the rationalization of newly identified ACs with single-atom modifications.

3. Results and discussion

3.1. Identification of activity cliffs produced by single-atom modifications

ACs with single-atom modifications capture minimal chemical changes leading to large potency variations and are thus of particular relevance for lead optimization efforts. Relevant single-atom modifications include specific atom replacements and positional changes of heteroatoms. Accordingly, we systematically searched bioactive compounds for atom-replacement or atom-walk APs. The resulting AP populations were then screened for ACs with an at least 100-fold difference in potency between paired analogs. The results are summarized in Table 1.

For both the IC_{50}- and K_i-based AP collections, it was found that 2.6%–3.1% of all APs formed ACs. Thus, large-magnitude potency changes as a consequence of single-atom modifications were

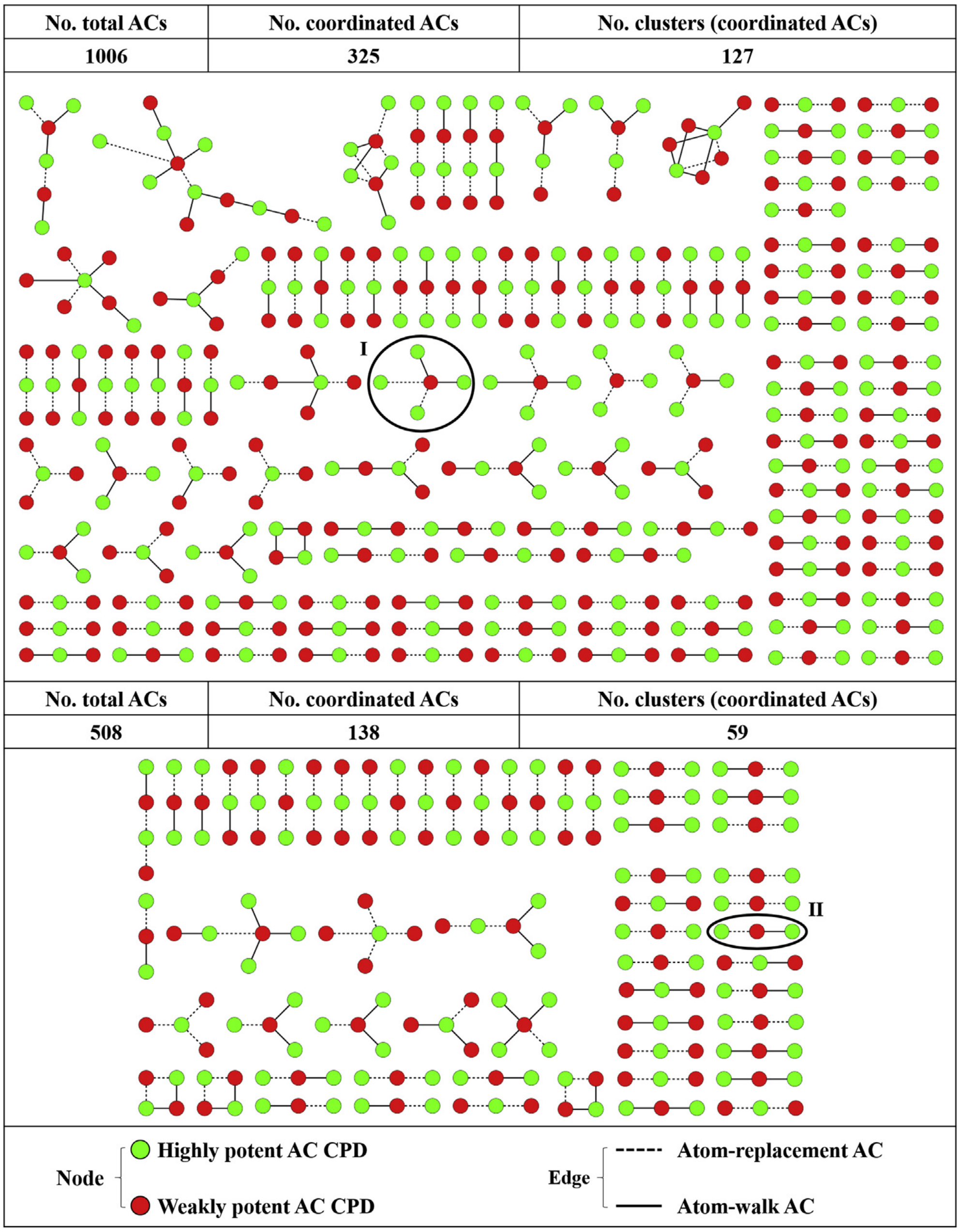


Fig. 2. Activity cliff networks. For the IC_{50} (top) and K_i (bottom) data set, AC networks with AC clusters formed by at least three compounds are shown. Nodes represent AC compounds and edges indicate the pairwise formation of ACs. Highly and weakly potent AC compounds are colored green and red, respectively. Dashed and solid black lines indicate atom-replacement and atom-walk ACs, respectively. Network statistics are provided. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

generally rare and less frequent than generally observed for analog pairs. On average, ~5% of exhaustively enumerated analog pairs meet or exceed a 100-fold difference in potency [3,6].

Given the very large number of qualifying APs we detected, in total more than 54,052, the small proportion of ~3% of ACs among them still resulted in large absolute numbers. The IC_{50} set yielded 722 atom-replacement and 284 atom-walk ACs. For the K_i set, the corresponding numbers were 384 and 124 ACs, respectively. Hence, atom-replacement ACs were more frequent than atom-walk ACs (proportional to the corresponding numbers of APs). In total, 1514 high-confidence ACs with single-atom modifications were obtained, a larger number than we anticipated, providing a substantial knowledge base for medicinal and computational chemistry applications. Fig. 1 shows exemplary atom-replacement and atom-walk ACs with activity against different targets.

3.2. Network analysis

To further investigate the formation of ACs with single-atom modifications, data set-dependent AC networks were generated, as shown in Fig. 2. These network views combine atom-replacement and atom-walk ACs and highlight the formation of coordinated ACs with three or more participating analogs. For the IC_{50} set, 325 of the total number of 1006 ACs (~32%) were formed in a coordinated manner, giving rise to 127 distinct AC clusters. For the K_i set, 138 of 508 ACs (~27%) were coordinated, resulting in 59

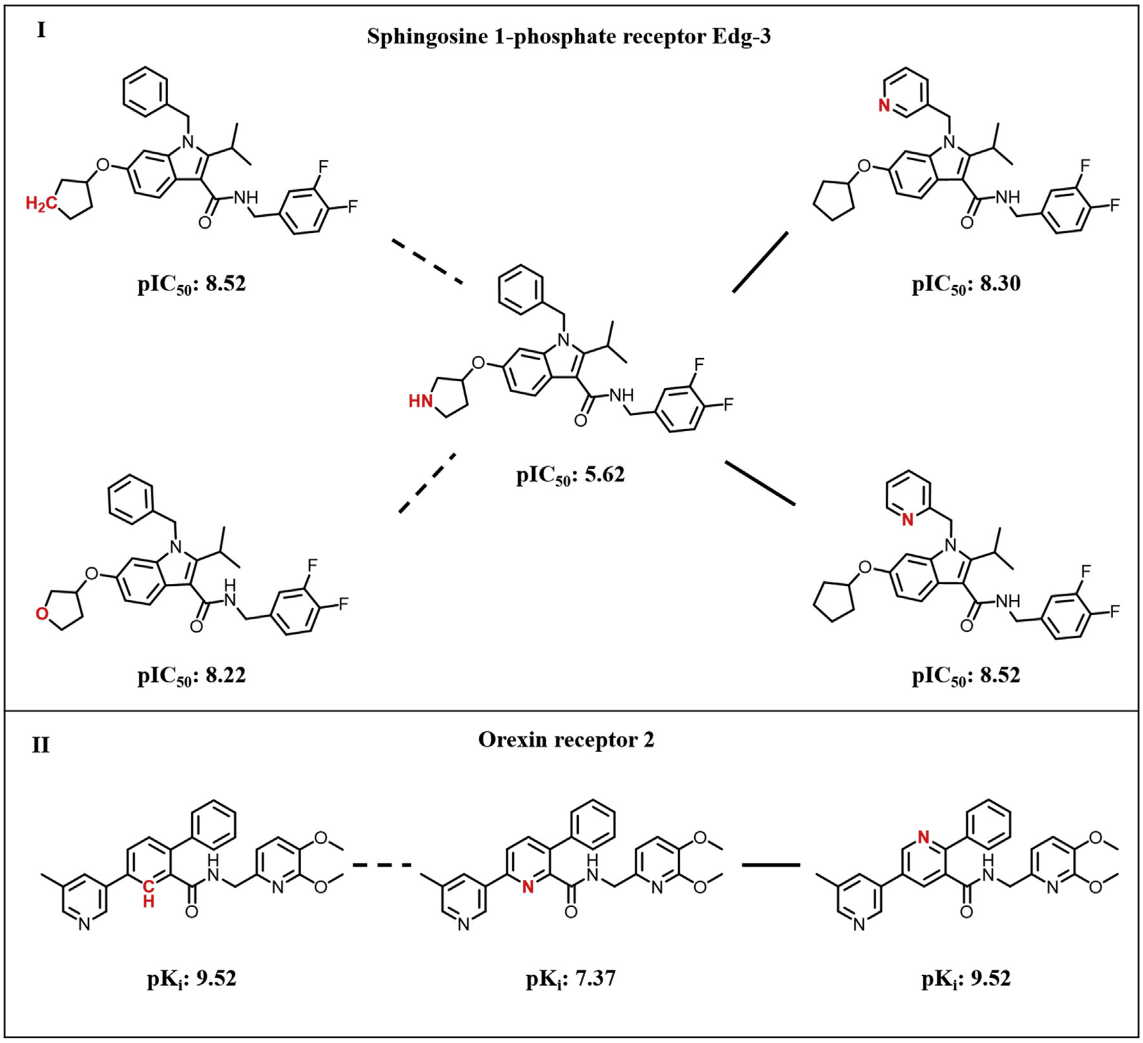


Fig. 3. Activity clusters. Two clusters highlighted in Fig. 2 (I and II) are shown in detail, which combine the formation of atom-replacement and atom-walk ACs. Target names and potency (pK_i or pIC_{50}) values are reported.

clusters. In both cases, the majority of clusters consisted of three compounds and only few large clusters with more complex topologies were observed. Thus, the majority of ACs with single-atom modifications was formed by individual compound pairs and only a limited number of AC compounds participated in more than one AC. These findings set ACs with single-atom modifications apart from previously studied ACs, more than 90% of which were formed in a coordinated manner [6]. For ACs with single-atom modifications, the much lower proportion of coordinated ACs was consistent with the lower than average rate of large potency variations as a consequence of subtle chemical changes, as discussed above. Accordingly, if such ACs are formed, they are likely to reveal key positions of SARs in analogs.

Network analysis also revealed that clusters of coordinated ACs with single-atom modifications often combined atom-replacement and atom-walk ACs. Hence, AC compounds frequently participated in both types of ACs. For example, in Fig. 2, two AC clusters are encircled (cluster I and II) that combine atom-replacement and atom-walk ACs in different ways. In Fig. 3, these clusters are depicted in detail. In both cases, a weakly potent AC compound formed different types of ACs with highly potent analogs, illustrating high SAR information content of these clusters.

3.3. Target distribution

Next, we determined the distribution of ACs with single-atom modifications across different target groups. The results are reported in Table 2 and Table 3 for the IC_{50} and K_i data set, respectively. In both cases, ACs were widely distributed across different classes of targets including a variety of enzymes, receptors, and ion channels. Overall, most ACs were available for G protein-coupled receptors (GPCRs) and protein kinases, which are among the most popular therapeutic targets. The IC_{50} set contained 310 ACs formed by kinase inhibitors and 147 ACs with ligands of family A GPCRs while the K_i set yielded 31 and 264 ACs formed by kinase inhibitors and family A GPCRs, respectively.

3.4. X-ray structures with activity cliff information

Given the subtle nature of ACs with single-atom modifications, we also intended to further explore possible reasons for AC formation. For studying ACs at the atomic level of detail, X-ray structures of ligand-target complexes provide a sound scientific basis. For GPCRs, structural data of ligand-target complexes are still sparse. By contrast, a wealth of structural information is currently available for kinases and their inhibitors. For all AC targets we

Table 2
Activity cliff distribution across different target groups (IC_{50} data set).

Target class	Target group	Number of ACs
Enzyme	Kinase	310
Membrane receptor	Family A G protein-coupled receptor	147
Enzyme	Protease	80
Enzyme	Transferase	72
Enzyme	Phosphodiesterase	71
Enzyme	Oxidoreductase	63
Ion channel	Voltage-gated ion channel	46
Enzyme	Hydrolase	37
Enzyme	Unclassified	31
Enzyme	Cytochrome P450	30
Transcription factor	Nuclear receptor	21
Transporter	Electrochemical transporter	19
Epigenetic regulator	Eraser	17
Membrane receptor	Family C G protein-coupled receptor	14
Other cytosolic protein	Unclassified	11
Enzyme	Ligase	7
Ion channel	Ligand-gated ion channel	7
Enzyme	Isomerase	5
Enzyme	Lyase	5
Epigenetic regulator	Writer	4
Ion channel	Other ion channel	2
Transcription factor	Unclassified	2
Membrane receptor	Unclassified	1
Membrane receptor	Family B G protein-coupled receptor	1
Secreted protein	Unclassified	1
Transporter	Primary active transporter	1
Unclassified protein	Unclassified	1

Table 3
Activity cliff distribution across different target groups (K_i data set).

Target class	Target group	No. ACs
Membrane receptor	Family A G protein-coupled receptor	264
Enzyme	Protease	58
Enzyme	Lyase	47
Enzyme	Kinase	31
Other cytosolic protein	Unclassified	16
Ion channel	Ligand-gated ion channel	15
Enzyme	Transferase	13
Transporter	Electrochemical transporter	8
Enzyme	Cytochrome P450	7
Enzyme	Hydrolase	7
Enzyme	Phosphodiesterase	6
Ion channel	Voltage-gated ion channel	6
Transcription factor	Nuclear receptor	6
Enzyme	Oxidoreductase	4
Epigenetic regulator	Eraser	4
Membrane receptor	Family B G protein-coupled receptor	4
Enzyme	Ligase	3
Epigenetic regulator	Reader	2
Membrane receptor	Family C G protein-coupled receptor	2
Unclassified protein	Unclassified	2
Enzyme	Unclassified	1
Enzyme	Isomerase	1
Ion channel	Other ion channel	1

identified, a systematic search was carried out for X-ray structures of complexes with AC compounds.

For the IC_{50} set, crystal structures of a lyase target in complex with both AC compounds were available, hence fully characterizing one AC. In addition, for 33 other ACs, a crystal structure of an AC target with one AC compound was identified, covering 10 different target classes. In 30 of 33 cases, the crystallographic AC compound represented the highly potent AC analog. Structures of kinase-inhibitor and GPCR-ligand complexes were associated with 15 and two ACs, respectively.

For the K_i set, two ACs were fully characterized by corresponding pairs of X-ray complexes and for 23 other ACs, an X-ray structure of an AC target with one AC compound was available, covering 10 target classes. In 22 of these cases, the crystallographic AC compound was the highly potent AC analog.

Taken together, for a total of 59 ACs with single-atom modifications, X-ray structures were available. Of these ACs, 50 involved atom replacements and the remaining nine examples represented atom walks, all of which involved nitrogen atoms. In addition, N—C was also the most frequent atom replacement, accounting for 29 of the 59 ACs.

3.5. Rationalization of activity cliffs

The presence or absence of specific ligand-target interactions as a consequence of compound modifications provides a possible rationale for AC formation. While contributions of specific interactions seen in X-ray structures to the free energy of binding must still be confirmed, ACs with single-atom modifications are particularly attractive for structure-based analysis because they frequently account for the presence or absence of atom-based interaction(s). Single-atom modifications as analyzed herein predominantly affect hydrogen bond formation or hydrogen-π interactions. The corresponding ACs thus provide an opportunity to focus on individual interactions that might implicated in large potency effects. Fig. 4 shows an exemplary AC from the K_i data set for which X-rays structures with both AC analogs were available. This nitrogen-walk AC was formed by inhibitors of coagulation factor X, a popular therapeutic protease target.

In this case, the “walking” nitrogen atom of the weakly potent AC analog forms a water-mediated hydrogen bond with factor X residue Ser214 that is absent in the highly potent AC analog when the nitrogen position changes. However, in the absence of this hydrogen bond, the highly potent analog forms a hydrogen bond with NH of residue Gly216 and another bifurcated water-mediated hydrogen bond involving the main chain carbonyl oxygen of residue Gly216 and the NH of residue Gly218. This rearranged hydrogen bonding pattern is accompanied by a conformational adjustment, as illustrated by the aligned binding modes of both AC

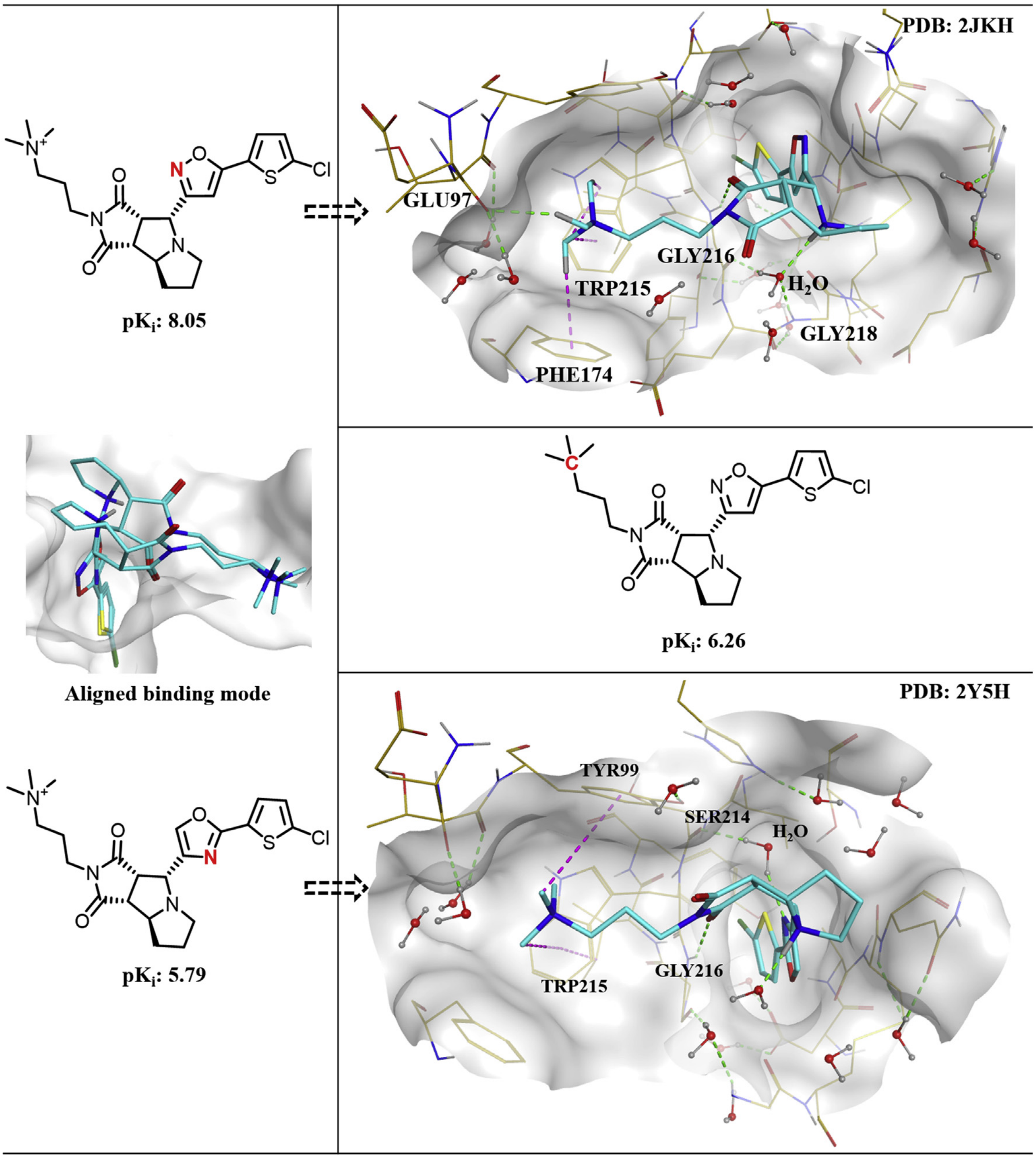


Fig. 4. X-ray structures representing an activity cliff. On the left, two factor X inhibitors are shown that form a nitrogen atom-walk AC and for which X-ray structures are available (right, with PDB IDs). Also shown is their binding mode alignment (obtained by superposition of the target structures in both complexes). A solvent-accessible surface view of the binding site is shown. Protein and ligand carbon atoms are colored yellow and cyan, respectively. Dashed green and magenta lines represent hydrogen bonds and hydrogen-π interactions, respectively. In the center separating the two X-ray structures, an analog of the AC compounds is shown that differs from the highly potent compound by a single N–C replacement. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

analogs, which partly alters their fit into the binding site.

Also shown is an inhibitor analog of these two AC compounds. This analog is only distinguished from the highly potent crystallographic AC compound by an N–C replacement, leading to the absence of the quaternary ammonium, which is involved in hydrogen bond and hydrogen-π interactions in the X-ray structure. Comparisons of these two analogs indicated that the N–C replacement and resulting loss of interactions involving the quaternary ammonium led to a nearly 100-fold loss in compound potency.

Surveying X-ray structures with AC compounds revealed a number of instances in which single-atom modifications represented by ACs could be associated with individual interactions, but also others where structural data did not provide a possible rationale for AC formation. Fig. 5 shows representative examples from different target classes.

Fig. 5a shows a nitrogen-walk AC formed by casein kinase II alpha inhibitors having a more than 1000-fold difference in potency. In the X-ray structure of the complex, the “walking” nitrogen forms a partly buried hydrogen bond with the NH of residue Val116, which could no longer be formed when the nitrogen changes its position. Whether or not this single interaction could account for the large potency differences would require further investigation.

In Fig. 5b, another large-magnitude nitrogen-walk AC is shown encoding a more than 1000-fold difference in potency. In this case, the respective nitrogen atom of the highly potent phosphodiesterase 10A inhibitor is involved in a network of hydrogen bonds that could not be formed in the case of the weakly potent analog. The absence of this hydrogen bond network is likely to cause a significant reduction in potency.

In the example in Fig. 5c, the N—O atom replacement converts a hydrogen bond donor of the highly potent factor X inhibitor into an acceptor in the weakly potent analog, which would affect two hydrogen bonds formed with different protein residues seen in the X-ray structure, which also provides a likely explanation for the observed reduction in potency.

The kinase inhibitors in Fig. 5d form a C—N atom-replacement AC. In the X-ray structure with the highly potent inhibitor, possible interactions involving the respective carbon atom become only apparent following a side chain rotamer adjustment of residue Phe49. In the presence of a preferred rotamer, a hydrogen-π interaction is possible. The C—N replacement captured by the AC would not permit this type of interaction in the case of the weakly potent analog.

The representative examples discussed so far illustrate different types of interaction hypotheses for rationalizing the formation of ACs with single-atom modifications. In some instance, well-defined interactions are very likely to be affected; in others, putative interaction differences between highly and weakly potent analogs are less evident. Finally, Fig. 5e presents an example where a

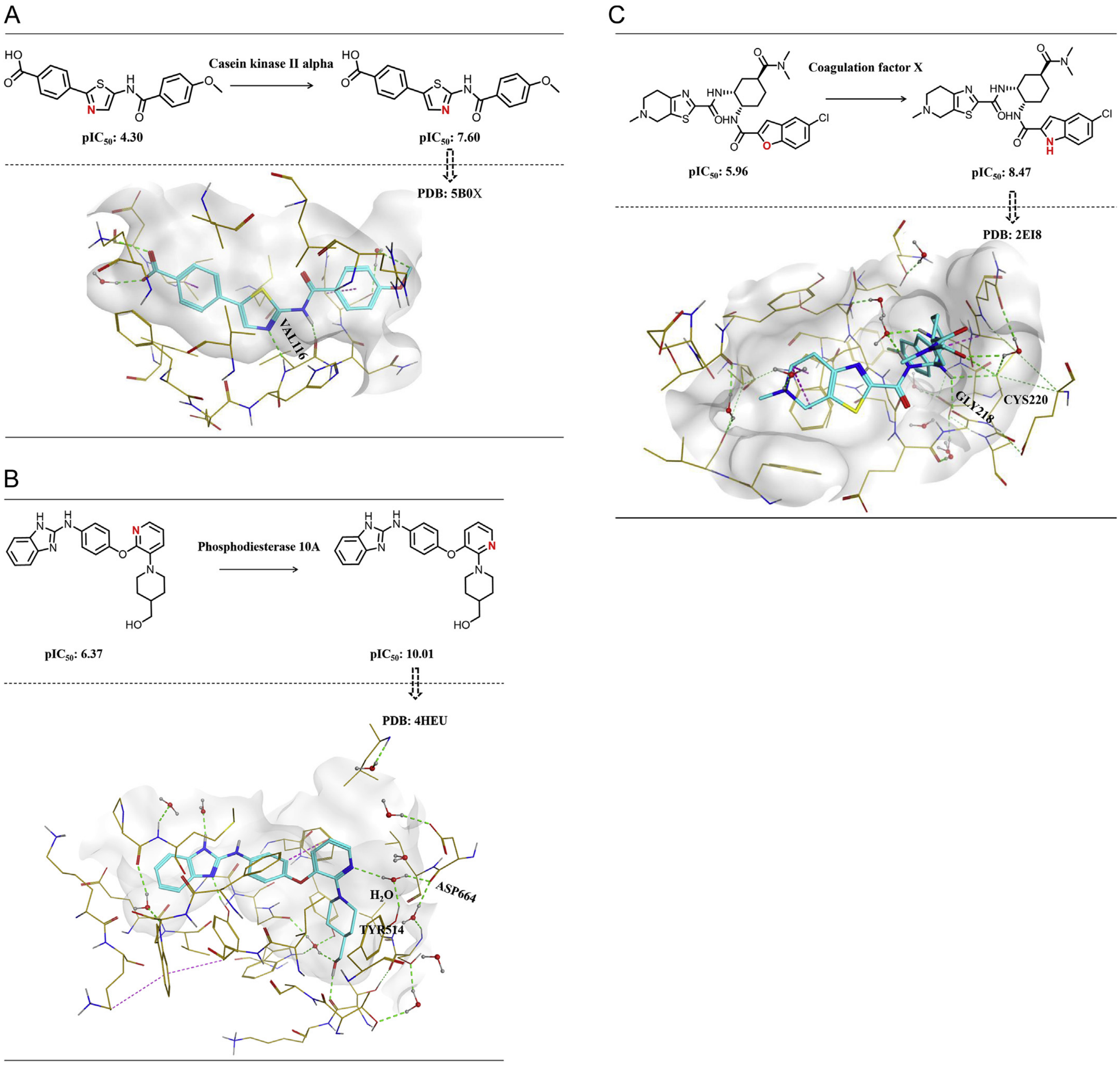


Fig. 5. Relating interactions in X-ray structures to activity cliffs. In **a**-**e**, exemplary ACs are shown for which an X-ray structure of the target with the highly potent (**a**–**d**) or weakly potent (**e**) AC analog was available (indicated by an arrow). The representation is according to Fig. 4.

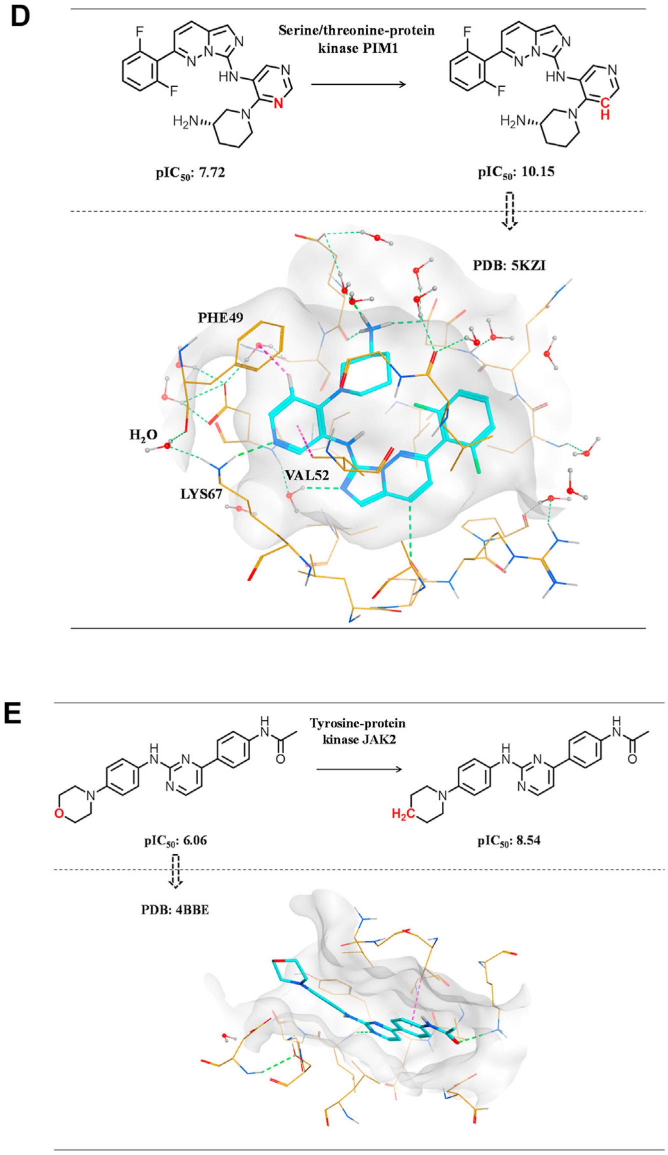


Fig. 5. (*continued*).

relevant X-ray structure does not provide a possible explanation for AC formation. In this case, an X-ray structure of the complex involving a weakly potent JAK2 kinase inhibitor was available that was converted into a highly potent analog through an O–C atom replacement. However, in the X-ray structure, the aliphatic ring in which the replacement occurred was exposed to the solvent environment at the entrance of the binding pocket and not involved in detectable interactions. Hence, in this case, it must be assumed that conformational changes accompanying interactions with the more hydrophobic analog and/or other components of the binding process such as entropic effects contributed to AC formation.

4. Conclusions

Herein, we have introduced a new category of ACs that capture single-atom modifications including atom replacements and walks. Thus, these ACs encode minimal chemical changes, which reflect corresponding compound optimization strategies, and reveal SAR information that can be attributed to interactions depending on individual atoms. We have systematically searched for these ACs and identified an unexpectedly large number of more than 1500 ACs with single-atom modifications and activity against a variety of targets. Network analysis showed that individual active compounds were capable of forming atom-replacement and atom-walk ACs with different analogs. Furthermore, we have searched for X-ray structures associated with these ACs that can be analyzed to study and potentially rationalize AC formation at the atomic level of detail. In addition, ligand-target interactions affected by single-atom modifications are likely to significantly contribute to SARs. The analysis revealed a variety of ways in which individual interactions might be affected, leading to large potency effects.

In summary, in this work, we have:

(1) introduced a new type of ACs with minimal chemical changes, i.e. single-atom replacements or atom walks;
(2) identified more than 1500 of these new ACs through large-scale activity data analysis;
(3) rationalized AC formation on the basis of X-ray structure of ligand-target complexes.

For SAR exploration, ACs capturing minimal chemical alterations provide a viable knowledge base. Furthermore, for computational chemistry and drug design, these ACs and associated X-ray structures provide interesting test cases because they make it possible to computationally probe individual interactions and their energetic contributions to binding. Therefore, we make our potency measurement-based collections of ACs with single-atom modifications and the associated structural information freely available in an open access deposition that is detailed in an associated Data in Brief note.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors thank OpenEye Scientific Software and Chemical Computing Group for academic software licenses. H.H. receives a PhD fellowship from the China Scholarship Council (CSC).

References

[1] C.G. Wermuth (Ed.), The Practice of Medicinal Chemistry, third ed., Academic Press, Burlington, MA, 2008.

[2] M. Wawer, E. Lounkine, A.M. Wassermann, J. Bajorath, Data structures and computational tools for the extraction of SAR information from large compound sets, Drug Discov. Today 15 (2010) 630–639.

[3] D. Stumpfe, Y. Hu, D. Dimova, J. Bajorath, Recent progress in understanding activity cliffs and their utility in medicinal chemistry, J. Med. Chem. 57 (2014) 18–28.

[4] G.M. Maggiora, On outliers and activity cliffs - why QSAR often disappoints, J. Chem. Inf. Model. 46 (2006), 1535-1535.

[5] J. Bajorath, Duality of activity cliffs in drug discovery, Expet Opin. Drug Discov. 14 (2019) 517–520.

[6] D. Stumpfe, H. Hu, J. Bajorath, Evolving concept of activity cliffs, ACS Omega 4 (2019) 14360–14368.

[7] H. Hu, J. Bajorath, Introducing a new category of activity cliffs combining different compound similarity criteria, RSC Med. Chem. 11 (2020) 132–141.

[8] S. Garai, P.M. Kulkarni, P.C. Schaffer, L.M. Leo, A.L. Brandt, A. Zagzoog, T. Black, X. Lin, D.P. Hurst, D.R. Janero, M.E. Abood, A. Zimmowitch, A. Straiker, R.G. Pertwee, M. Kelly, A.-M. Szczesniak, E.M. Denovan-Wright, K. Mackie, A.G. Hohmann, P.H. Reggio, R.B. Laprairie, G.A. Thakur, Application of fluorine- and nitrogen-walk approaches: defining the structural and functional diversity of 2-phenylindole class of cannabinoid 1 receptor positive allosteric modulators, J. Med. Chem. 63 (2020) 542–568.

[9] L.D. Pennington, B.M. Aquila, Y. Choi, R.A. Valiulin, I. Muegge, Positional analogue scanning: an effective strategy for multiparameter optimization in drug design, J. Med. Chem. 63 (2020) 8956–8976.

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

[11] RDKit, Cheminformatics and Machine Learning Software, 2013. http://www.rdkit.org.

[12] D. Stumpfe, D. Dimova, J. Bajorath, Composition and topology of activity cliff clusters formed by bioactive compounds, J. Chem. Inf. Model. 54 (2014) 451–461.

[13] M.E. Smoot, K. Ono, J. Ruscheinski, P.L. Wang, T. Ideker, Cytoscape 2.8: new features for data integration and network visualization, Bioinformatics 27 (2011) 431–432.

[14] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein Data Bank, Nucleic Acids Res. 28 (2000) 235–242.

[15] UniProt Consortium, UniProt: a hub for protein information, Nucleic Acids Res. 43 (2015) D204–D212.

[16] M.R. Berthold, N. Cebron, F. Dill, T.R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, B. Wiswedel, KNIME: the konstanz information miner, in: C. Preisach, H. Burkhart, L. Schmidt Thieme, R. Decker (Eds.), Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin, 2008, pp. 319–326.

[17] S. Vilar, G. Cozza, S. Moro, Medicinal chemistry and the molecular operating environment (moe): application of QSAR and molecular docking to drug discovery, Curr. Top. Med. Chem. 8 (2008) 1555–1572.

[18] Molecular Operating Environment (MOE), 01; Chemical Computing Group ULC, 1010 Sherbooke St. West, 2018. Suite #910, Montreal, Canada, H3A 2R7, 2018.

Summary

In this study, more than 1500 ACs with single-atom modifications were identified. This AC category is generally rare accounting for around 3% of qualifying compound pairs. The majority of them (1106 ACs) were atom-replacement ACs. An AC network demonstrated that $\sim$ 30% of all ACs were formed in a coordinated manner. Typically, this AC category was frequently observed for protein kinases and G protein-coupled receptors, but ACs for other target families were also detected.

For these ACs, the crystallographic complexes for both cliff compounds were only available in three cases. In addition, we also detected 56 ACs for which only one X-ray ligand-target complex was available. Taken together, these 59 ACs provided the possibility to rationalize AC formation based on crystallographic structures. Obvious interaction differences such as hydrogen bond formation or hydrogen-π interactions between AC-forming compounds were observed which could be directly attributed to individual heteroatom modifications. Since the X-ray structures only provide an incomplete picture of binding events, further experimental tests are required to probe the binding free energy contributions of particular interactions. The identified ACs with single-atom modifications have been made available to the public for follow-up studies.

The formation of ACs is generally a rare event. Currently, ACs are often globally analyzed and the criterion of at least two orders of magnitude difference in potency is universally applied irrespective of compound activity classes. In the next chapter, we attempt to rationalize AC formation across different activity classes by relating potency value distribution and structural similarity relationships to each other.

Chapter 7

Rationalizing the Formation of Activity Cliffs in Different Compound Data Sets

Introduction

Primary AC analysis has been mainly focused on globally exploring AC characteristics such as coordinated or isolated ACs, frequency of occurrence, or extraction of associated SAR information from AC clusters. Large-scale AC studies indicate that the formation of ACs is a rare event accounting for only $\sim$ 5% of structurally similar compound pairs. However, the reason behind this statistic has not yet been explored in detail. Compound potency value distributions in activity classes often significantly differ. Thus, the analysis of AC characters on a per activity class (target) basis could provide further insight into its AC formation.

In this chapter, activity classes with high-confidence activity data from the ChEMBL database were systematically extracted. The qualifying activity classes were assigned to different categories according to their potency value distributions. RMMP-based networks were constructed to visualize and rationalize the difference in AC formation across different activity classes.

ACS OMEGA

Article

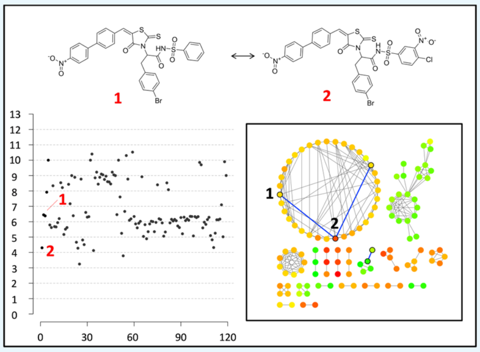
Cite This: ACS Omega 2018, 3, 7736−7744

Rationalizing the Formation of Activity Cliffs in Different Compound Data Sets

Huabin Hu, Dagmar Stumpfe, and Jürgen Bajorath*

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

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



1. INTRODUCTION

Activity cliffs are formed by structurally similar (analogous) active compounds with large differences in potency.[1−4] Because activity cliffs represent small chemical changes having large biological activity effects, they embody the pinnacle of structure−activity relationship (SAR) discontinuity,[3] which is detrimental for quantitative SAR predictions.[2] However, discontinuous SARs and activity cliffs often reveal SAR determinants, especially when encountered during early stages of compound optimization, and thus provide viable information for medicinal chemistry.[3,4]

For a consistent assessment of activity cliffs, similarity and potency difference criteria must be clearly defined.[3] On the basis of globally assessed potency range distributions of pairs of active analogues, an at least 100-fold difference in potency (on the basis of equilibrium constants, if available) has been proposed and frequently been used as an activity cliff criterion.[4,5] The definition of activity cliffs also depends on the molecular representations and similarity measures that are used.[4,6] Compound similarity for activity cliff definition can be quantified in different ways, for example, by calculating Tanimoto similarity on the basis of molecular fingerprint representations or by applying substructure-based similarity criteria.[3,4] Numerical similarity measures, such as the Tanimoto coefficient, yield a continuum of values, and a threshold must be set for defining activity cliffs. By contrast, substructure-based methods produce a binary readout, for example, two compounds share the same core structure—and are classified as similar—or they do not. In addition to comparing molecular graph-based (two-dimensional) representations, activity cliffs have also been determined in three dimensions by calculating the similarity of experimental compound binding modes taken from complex X-ray structures.[7]

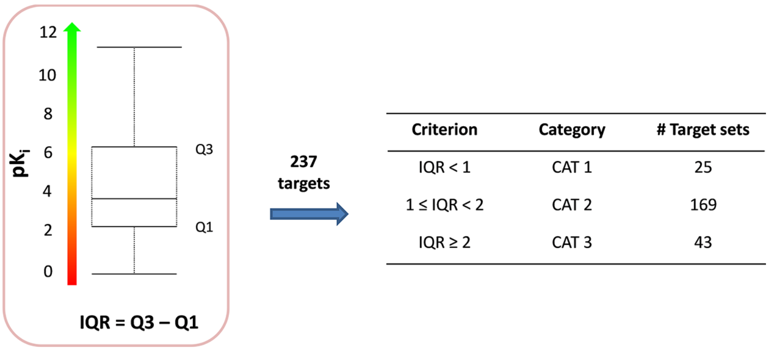
For graph-based activity cliff definition, substructure similarity assessment is—in our experience—generally more consistent than numerical similarity calculations and often easier to interpret from a chemical perspective.[4] Among substructure-based approaches, the matched molecular pair (MMP) concept[8,9] is particularly attractive for activity cliff definition. An MMP is defined as a pair of compounds that are only distinguished by a chemical modification at a single site.[8] This modification corresponds to the exchange of a pair of substructures,[8,9] which is termed a chemical transformation.[9] By introducing appropriate transformation size restrictions, the formation of MMPs can be limited to structural analogues typically generated during compound optimization.[10] Applying this similarity criterion yields a structurally conservative and chemically intuitive definition of activity cliffs.[4,10] Moreover, transformation size-restricted MMPs can be efficiently generated algorithmically,[9,10] hence enabling large-scale analysis of activity cliff populations.

In light of these considerations, our preferred activity cliff definition encompasses the formation of a transformation size-restricted MMP by two compounds sharing the same biological activity that have an at least 100-fold difference in potency.[4,10] Whenever possible, potency differences are determined on the basis of (assay-independent) equilibrium constants. The so-defined activity cliffs have been termed MMP-cliffs.[10]

The definition of activity cliffs is focused on compound pairs and hence accounts for pairwise relationships. However, activity cliffs in compound data sets are mostly not formed by isolated compound pairs (i.e., pairs without structural neighbors forming

Received: May 30, 2018
Accepted: June 26, 2018
Published: July 11, 2018

DOI: 10.1021/acsomega.8b01188
ACS Omega 2018, 3, 7736−7744



Criterion	Category	# Target sets
IQR < 1	CAT 1	25
1 ≤ IQR < 2	CAT 2	169
IQR ≥ 2	CAT 3	43

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

additional activity cliffs). Rather, the vast majority of activity cliffs are formed in a coordinated manner by groups of structurally related compounds with large potency variations, meaning that individual compounds are involved in the formation of multiple activity cliffs with different analogues.[11,12] In activity cliff networks where nodes represent compounds and edges pairwise activity cliffs, compound subsets forming coordinated cliffs give rise to the formation of disjoint clusters.[12] These activity cliff clusters are a rich source of SAR information and much more informative than cliffs considered as isolated.[13] More than 95% of MMP-cliffs detected across different data sets were formed in a coordinated manner.[14] In activity cliff networks, clusters often include "hubs," that are, nodes representing molecules that are centers of local activity cliff formation with multiple partner compounds. Such molecules have also been termed "activity cliff generators."[15,16]

In addition to activity cliff coordination, the frequency with which activity cliffs occur across different data sets has been determined.[5,14] There has been substantial growth in activity cliff information over time. For example, from June 2011 until January 2015, the number of MMP-cliffs originating from the ChEMBL database,[17] the major public repository of compounds and activity data from medicinal chemistry sources, nearly doubled; with a total of more than 17 000 MMP-cliffs available at the beginning of 2015.[14] In addition, the target coverage of MMP-cliffs increased from about 200 to 300 individual target proteins over this period of time. However, despite this strong growth, the proportion of bioactive compounds involved in the formation of MMP-cliffs across different compound data sets remained essentially constant at close to 23%.[14]

So far, activity cliffs have been studied in exemplary compound sets on a case-by-case basis or surveyed by global statistical analysis.[5,14] In addition, cliff populations have been organized and visualized in network representations.[12,13] However, what has not been attempted thus far is systematically exploring and comparing activity cliff formation in different compound activity classes (also called target sets). To these ends, we have analyzed in detail potency distributions and structural relationships between compounds in many different target sets, studied how activity cliffs were formed, and determined the differences between sets. Hence, the focus of our current study has been on details of activity cliff arrangements in individual compound sets rather than on global statistical exploration. Our analysis revealed many characteristic differences in activity cliff formation between target sets.

2. MATERIALS AND METHODS

2.1. Activity Cliff Definition. For our current analysis, we introduced a modification of our preferred MMP-cliff definition stated above.[4,10] For MMP generation, standard random fragmentation of exocyclic single bonds[9] was replaced by fragmentation according to retrosynthetic (RECAP) rules,[18] yielding (transformation size-restricted) RECAP-MMPs (RMMPs).[19] Retrosynthetic MMPs were generated to further increase the chemical relevance (synthetic accessibility) of compound pairs, forming cliffs. Accordingly, the formation of an RMMP was used as a similarity criterion for activity cliffs, and an at least 100-fold difference in potency between RMMP compounds was required, as before. The so-defined activity cliffs are referred to as RMMP-cliffs.

2.2. Compounds and Activity Data. Bioactive compounds with high-confidence activity data were assembled from ChEMBL version 23.[17] The following selection criteria were applied: First, only compounds involved in direct interactions (type "D") with human targets at the highest confidence level (assay confidence score 9) were selected. Second, only numerically specified equilibrium constants (K_i values) were considered as potency measurements. Equilibrium constants were reported as pK_i values. On the basis of these selection criteria, a total of 71 967 unique compounds were obtained with activity against a total of 904 targets. Accordingly, these compounds were organized into 904 target sets.

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

For each target set, an RMMP network was generated in which nodes represented compounds and edges pairwise RMMP relationships. In this network, each separate RMMP cluster represented a unique series of analogues. RMMP networks were also used to represent RMMP-cliffs by highlighting edges that represented both RMMP and activity cliff relationships. All network representations were drawn with Cytoscape.[20]

2.4. Potency Distributions. For the 237 qualifying target sets, compound potency distributions were monitored in boxplots. On the basis of the interquartile range (IQR), that is, the range between quartile 1 (Q1) and 3 (Q3), target sets

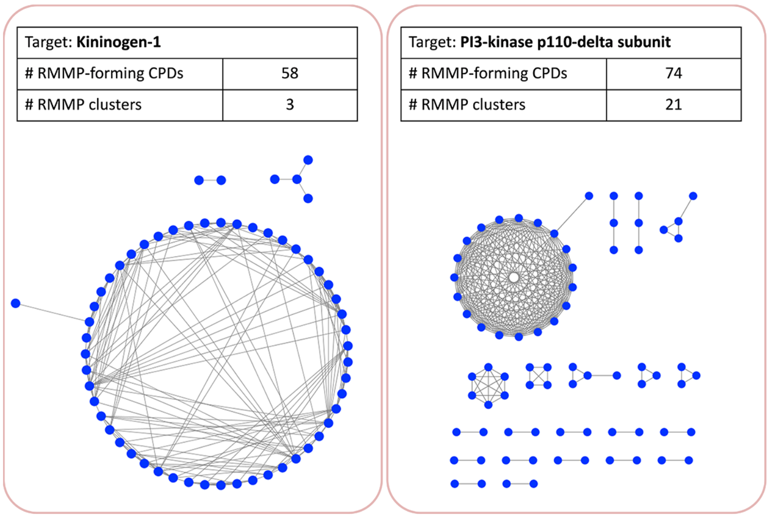


Figure 2. Structural similarity in target sets. For two exemplary target sets, RMMP networks are shown in which blue nodes represent compounds and edges pairwise RMMP relationships. Separate clusters represent a unique series of analogues. Although the number of RMMP-forming compounds (CPDs) was similar for both target sets, the number of clusters differed significantly.

were assigned to three different categories, as shown in Figure 1: category 1 (CAT 1), IQR was smaller than 1 order of magnitude (<10-fold difference in potency); CAT 2, IQR fell between 1 and less than 2 orders of magnitude (10- to 100-fold difference); and CAT 3, IQR no smaller than 2 orders of magnitude (≥100-fold difference in potency). By definition, the IQR represented the potency range of ~50% of the compounds in each target set.

3. RESULTS AND DISCUSSION

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

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

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

3.4. RMMP-Cliffs. In 207 of the 237 qualifying targets sets, RMMP-cliffs were identified, amounting to a total of 11 834 cliffs. Table 1 reports that the number of RMMP-cliffs increased over target sets of CAT 1, 2, and 3, with on average 2, 52, and 69 cliffs per set, respectively. Thus, there was a general trend of increasing number of RMMP-cliffs with increasing variability of potency distributions. The very small number of RMMP-cliffs for CAT 1 sets was directly attributable to the narrow potency distributions characterizing this category. Table 2 reports that the 48 target sets containing 50 to a maximum of 820 RMMP-cliffs exclusively belonged to CAT 2 and CAT 3 that had activity cliff-relevant IQR values. By contrast, target sets with less than 50 RMMP-cliffs were found in all 3 categories. Figure 4 shows that the majority of target sets with large number of 100 or more

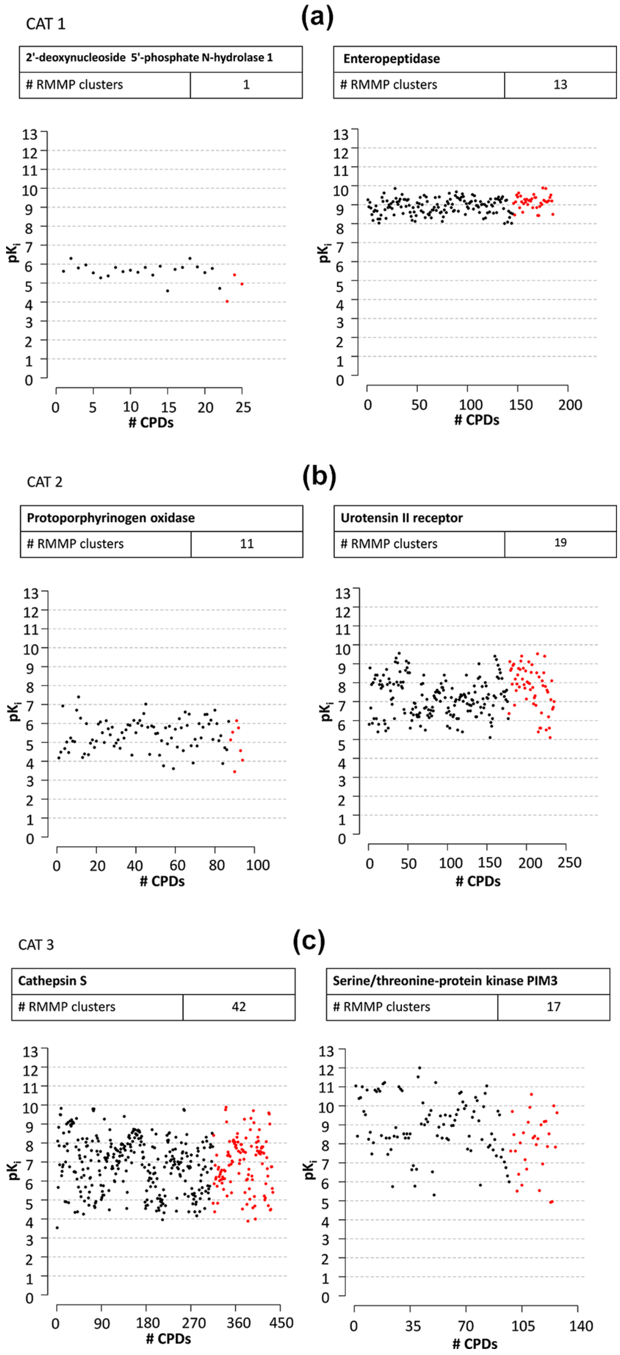


Figure 3. Potency distribution profiles. Shown are exemplary potency distribution profiles for target sets belonging to different categories [(a), CAT 1; (b), CAT 2; (c), CAT 3] according to Figure 1. Black dots represent RMMP compounds and red dots singletons not participating in RMMPs.

Table 1. Target Set Statistics[a]

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

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

Table 2. RMMP-Cliff Distribution[a]

# RMMP-cliffs (range)	# target sets	CATs
0	30	1, 2, 3
[1, 10)	77	1, 2, 3
[10, 20)	33	1, 2, 3
[20, 50)	49	1, 2, 3
[50, 100)	20	2, 3
[100, 500)	25	2, 3
[500, 820]	3	2, 3

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

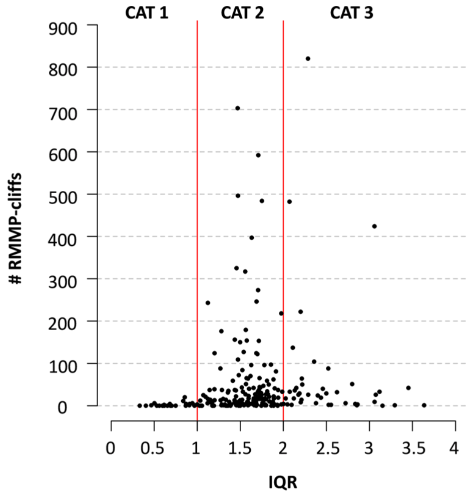
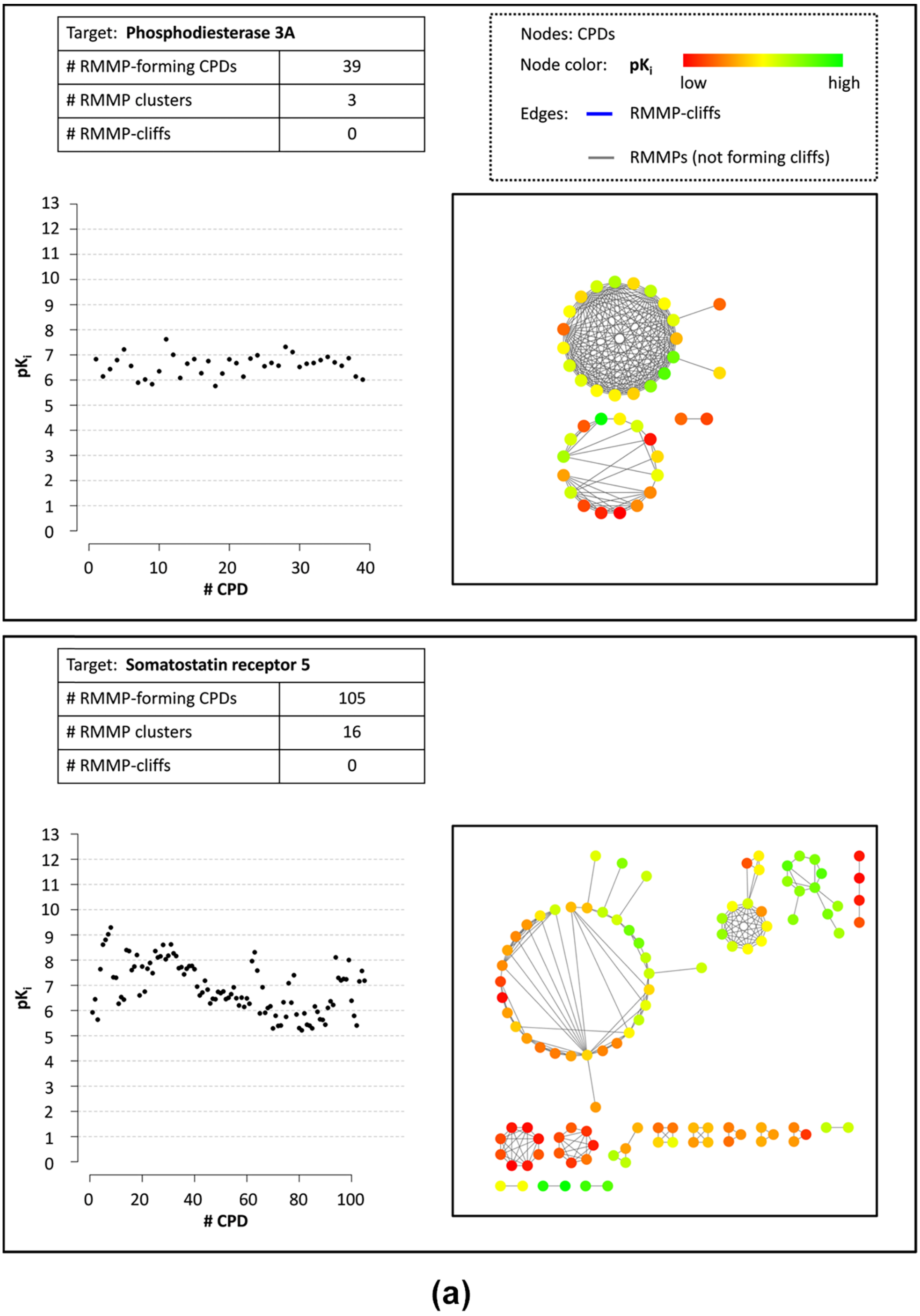


Figure 4. RMMP-cliffs vs IQR values. For each of the 237 target sets, the number of RMMP-cliffs (y-axis) is plotted against increasing IQR values (x-axis). Red vertical lines separate target sets belonging to CAT 1, 2, and 3.

RMMP-cliffs belonged to CAT 2, which was due to the large number of 169 target sets in this category compared to only 43 sets in CAT 3. A systematic increase in the number of activity cliffs with increasing IQR values was not observed.

However, despite these general trends, the propensity to form RMMP-cliffs could not solely be attributed to the variability and spread of potency distributions. Rather, as further discussed below, potency distributions in target sets must be viewed in combination with RMMP networks and their cluster structure. Table 1 also reports that target sets in CAT 1, 2, and 3 contained on average 10, 54, and 37 RMMP clusters, respectively. Thus, CAT 2 and CAT 3 sets contained large number of clusters (analogue series) whose local potency distributions strongly influenced RMMP-cliff formation.

3.5. Interplay of Potency Patterns and Structural Relationships. The 207 target sets containing RMMP-cliffs were individually examined to evaluate potency distribution profiles and RMMP networks in context and rationalize why RMMP-cliffs were formed with different frequencies. The analysis revealed a number of characteristic features determining cliff formation that are summarized in Figure 5 by comparing exemplary target sets. Figure 5a (top) shows a set of phosphodiesterase 3A inhibitors with a flat CAT 1 potency distribution profile, which prohibited RMMP-cliff formation, despite the presence of two analogue series with in-part

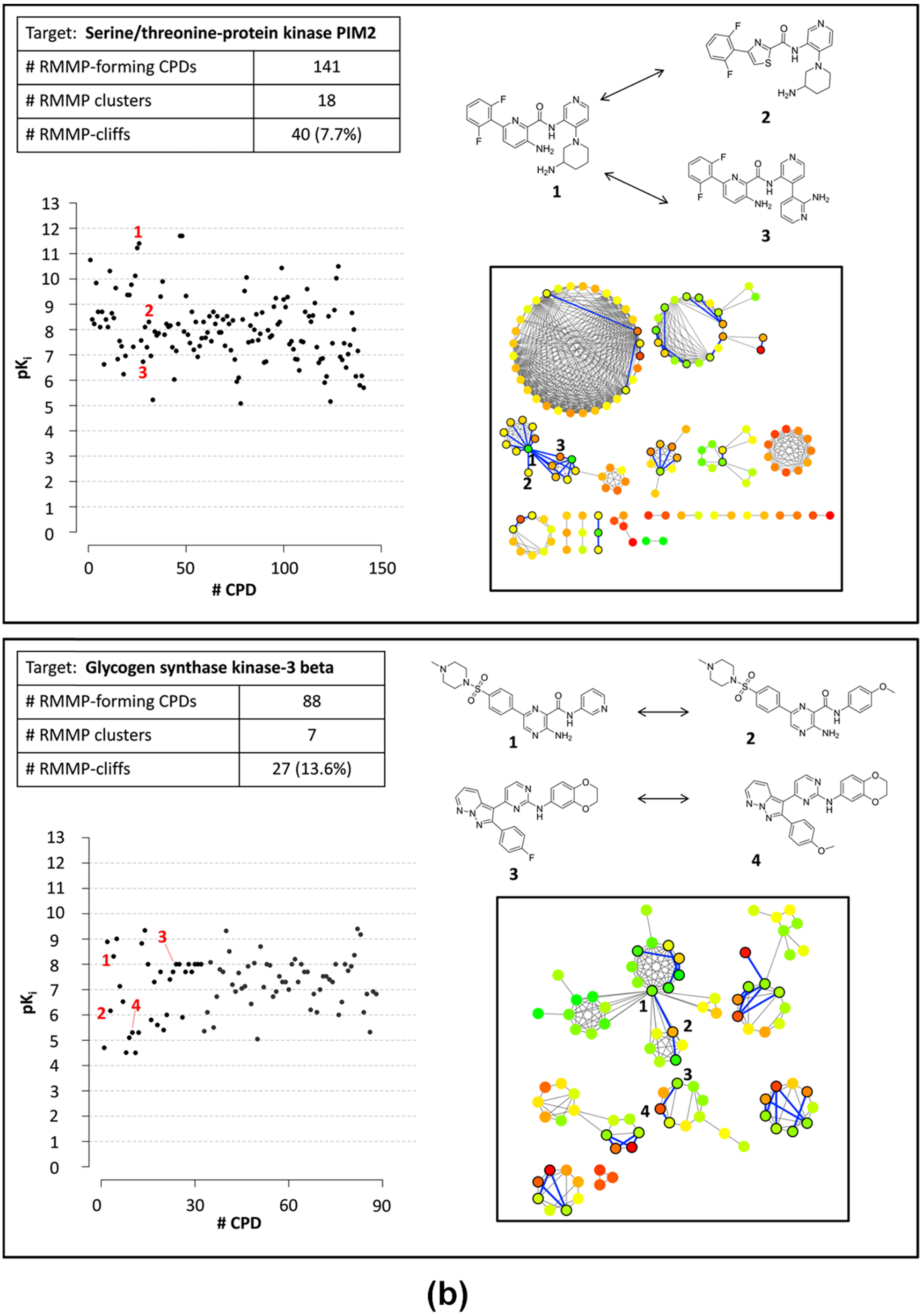


Target: **Phosphodiesterase 3A**	
# RMMP-forming CPDs	39
# RMMP clusters	3
# RMMP-cliffs	0

Target: **Somatostatin receptor 5**	
# RMMP-forming CPDs	105
# RMMP clusters	16
# RMMP-cliffs	0

Figure 5. continued

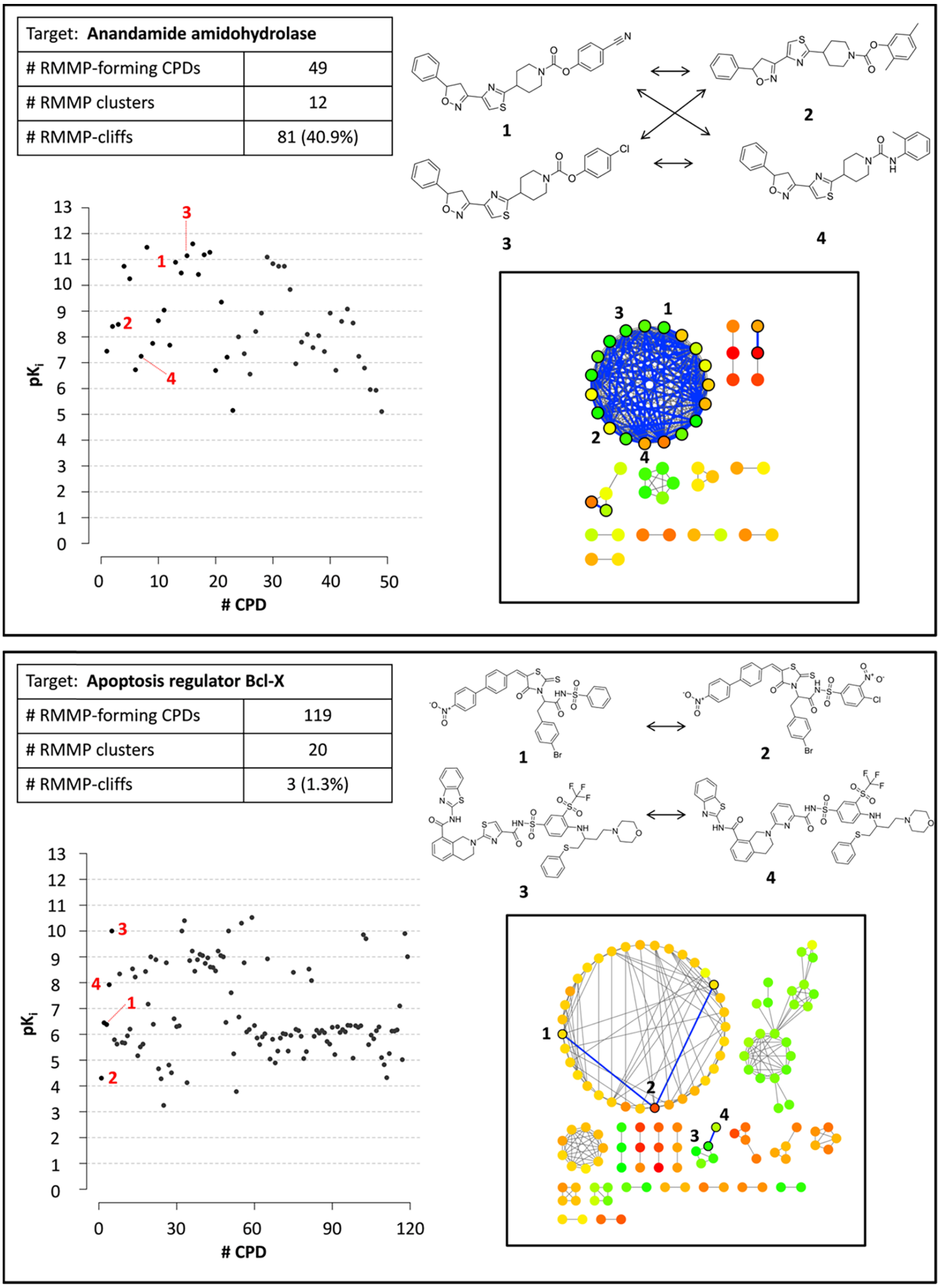
DOI: 10.1021/acsomega.8b01188



Target: Serine/threonine-protein kinase PIM2	
# RMMP-forming CPDs	141
# RMMP clusters	18
# RMMP-cliffs	40 (7.7%)

Target: Glycogen synthase kinase-3 beta	
# RMMP-forming CPDs	88
# RMMP clusters	7
# RMMP-cliffs	27 (13.6%)

Figure 5. continued



Target: **Anandamide amidohydrolase**	
# RMMP-forming CPDs	49
# RMMP clusters	12
# RMMP-cliffs	81 (40.9%)

Target: **Apoptosis regulator Bcl-X**	
# RMMP-forming CPDs	119
# RMMP clusters	20
# RMMP-cliffs	3 (1.3%)

(c)

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

extensive RMMP relationships. In addition, Figure 5a (bottom) displays somatostatin receptor 5 ligands with a variable CAT 2 distribution and more than 100 RMMP-forming compounds. Although cliff formation was more likely in this case, the target set did not contain any RMMP-cliffs either. This was a direct consequence of a heterogeneous cluster structure and local potency distributions over different subsets of analogues forming 16 clusters, as revealed by the RMMP network of this set.

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

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

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

4. CONCLUSIONS

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

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

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: bajorath@bit.uni-bonn.de. Phone: 49-228-73-69100.

ORCID

Jürgen Bajorath: 0000-0002-0557-5714

Author Contributions

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

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

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

■ REFERENCES

(1) Lajiness, M. Evaluation of the Performance of Dissimilarity Selection Methodology. In *QSAR: Rational Approaches to the Design of Bioactive Compounds*; Silipo, C., Vittoria, A., Eds.; Elsevier: Amsterdam, Netherlands, 1991; pp 201−204.

(2) Maggiora, G. M. On Outliers and Activity Cliffs − Why QSAR often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535.

(3) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J. Med. Chem.* **2012**, *55*, 2932−2942.

(4) Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57*, 18−28.

(5) Stumpfe, D.; Bajorath, J. Frequency of Occurrence and Potency Range Distribution of Activity Cliffs in Bioactive Compounds. *J. Chem. Inf. Model.* **2012**, *52*, 2348−2353.

(6) Medina-Franco, J. L.; Martínez-Mayorga, K.; Bender, A.; Marín, R. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. Characterization of Activity Landscapes using 2D and 3D Similarity Methods: Consensus Activity Cliffs. *J. Chem. Inf. Model.* **2009**, *49*, 477−491.

(7) Hu, Y.; Furtmann, N.; Gütschow, M.; Bajorath, J. Systematic Identification and Classification of Three-Dimensional Activity Cliffs. *J. Chem. Inf. Model.* **2012**, *52*, 1490−1498.

(8) Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2005; pp 271−285.

(9) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 339−348.

(10) Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J. MMP-Cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs. *J. Chem. Inf. Model.* **2012**, *52*, 1138−1145.

(11) Vogt, M.; Huang, Y.; Bajorath, J. From Activity Cliffs to Activity Ridges: Informative Data Structures for SAR Analysis. *J. Chem. Inf. Model.* **2011**, *51*, 1848−1856.

(12) Stumpfe, D.; Dimova, D.; Bajorath, J. Composition and Topology of Activity Cliff Clusters Formed by Bioactive Compounds. *J. Chem. Inf. Model.* **2014**, *54*, 451−461.

(13) Dimova, D.; Stumpfe, D.; Hu, Y.; Bajorath, J. Activity Cliff Clusters as a Source of Structure-Activity Relationship Information. *Expert Opin. Drug Discovery* **2015**, *10*, 441−447.

(14) Stumpfe, D.; Bajorath, J. Monitoring Global Growth of Activity Cliff Information over Time and Assessing Activity Cliff Frequencies and Distributions. *Future Med. Chem.* **2015**, *7*, 1565−1579.

(15) Méndez-Lucio, O.; Pérez-Villanueva, J.; Castillo, R.; Medina-Franco, J. L. Identifying Activity Cliff Generators of PPAR Ligands Using SAS Maps. *Mol. Inf.* **2012**, *31*, 837−846.

(16) Pérez-Villanueva, J.; Méndez-Lucio, O.; Soria-Arteche, O.; Medina-Franco, J. L. Activity Cliffs and Activity Cliff Generators Based on Chemotype-Related Activity Landscapes. *Mol. Diversity* **2015**, *19*, 1021−1035.

(17) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.

(18) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP-Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful

DOI: 10.1021/acsomega.8b01188

Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511−522.

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

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

Summary

In this study, the differences in the frequency of occurrence of AC formation across activity classes were investigated by relating potency distribution and structural similarity to each other. To obtain statistically sound results, we only selected activity classes containing at least 100 RMMPs. These qualifying activity classes were assigned to three categories according to the interquartile range (IQR) values of compound potency distributions monitored in boxplots. IQR values indicated the likelihood of the activity classes meeting the criterion of a large difference in potency. For the majority of activity classes, the IQR values fell between one and two orders of magnitude. However, increasing IQR values did not correlate with increasing number of ACs. Given a large potency variation, structural similarity has to be taken into account to evaluate the possibility of AC formation. To this end, a RMMP-based network was constructed where edges highlighted the AC relationships.

The study demonstrated that the likelihood of AC formation largely depended on the potency variations of intra-RMMP-clusters. For some activity classes with large potency fluctuations, if intra-cluster potency levels were comparable, the formation of ACs was less likely. Conversely, a proportion of activity classes had high propensity for AC formation, which was a consequence of the structural homogeneity and large potency variation within the clusters. The integration of potency value distribution and RMMP-based network analysis made it possible to rationalize the differences of AC frequency among activity classes.

Since potency distributions across activity classes often differ significantly, universal application of the “at least two orders of magnitude difference in potency" criterion might not be appropriate. In the next chapter, we derive statistically determined potency difference criteria tailored for individual activity classes.

Chapter 8

Second-Generation Activity Cliffs Identified on the Basis of Target Set-Dependent Potency Difference Criteria

Introduction

Many efforts in identifying ACs focused on different molecular representations to assess molecular similarity, yielding different AC categories such as (R)MMP-cliffs, fingerprint-based cliffs or 3D-cliffs. Less attention has been paid to studying the potency difference criteria for AC assessment. Typically, a general threshold of at least 100-fold difference in potency is frequently applied.

As the results of the preceding chapter showed, the interplay of potency variations and structural similarity relationships within the RMMP-based clusters strongly influenced the frequency of occurrence of ACs. In this chapter, we derived an activity class-dependent potency difference criterion for AC assessment.

Reprinted with permission from “Hu, H.; Stumpfe, D.; Bajorath, J. Second-Generation Activity Cliffs Identified on the Basis of Target Set-Dependent Potency Difference Criteria. *Future Medicinal Chemistry* **2019**, *33*, 559-572”. *Copyright 2019 Future Science Group*

Research Article

For reprint orders, please contact: reprints@future-science.com

Future Medicinal Chemistry

Second-generation activity cliffs identified on the basis of target set-dependent potency difference criteria

Huabin Hu[1], Dagmar Stumpfe[1] & Jürgen Bajorath*,[1]

[1]Department of Life Science Informatics, b-it, LIMES Program Unit Chemical Biology & Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Endenicher Allee 19c, D-53115 Bonn, Germany

*Author for correspondence: Tel.: +49 228 7369 100; Fax: +49 228 7369 101; bajorath@bit.uni-bonn.de

Aim: Activity cliffs (ACs) are formed by structurally similar compounds with large potency differences. Accordingly, ACs reveal determinants of structure–activity relationships. This makes ACs highly interesting and relevant for medicinal chemistry and chemoinformatics. So far, ACs have been defined on the basis of generally applied molecular similarity and potency difference criteria. **Results:** We present the first assessment of ACs taking target set-dependent compound potency distributions into account, leading to a new target set-dependent definition of ACs. The formation of these ACs is analyzed in detail. **Conclusion:** Second-generation ACs are obtained on the basis of target set-dependent potency difference thresholds. Compared with generally defined ACs, target set-dependent ACs have further increased medicinal chemistry relevance.

First draft submitted: 5 July 2018; Accepted for publication: 6 December 2018; Published online: 19 March 2019

Keywords: activity cliffs • molecular similarity • potency differences • potency value distributions • structure–activity relationships • target set-dependent activity cliff definition

In medicinal chemistry and chemoinformatics, a pair of structurally similar compounds with a significant potency difference constitutes an activity cliff (AC) [1–3]. ACs are of particular interest for the analysis of structure–activity relationships (SARs) and chemical optimization because they encode small structural changes with large potency effects [2]. To evaluate ACs in a consistent manner and systematically analyze ACs in different target sets (compound activity classes), two criteria must be clearly defined: the 'similarity' criterion and the 'potency difference' criterion [2–4].

For ACs, molecular similarity has originally been calculated on the basis of fingerprint descriptors and the Tanimoto similarity metric [1,2]. Setting thresholds of calculated similarity values for AC formation is convenient for computational analysis, but calculated similarity relationships are often difficult to reconcile in medicinal chemistry terms. Accordingly, molecular similarity has also been assessed on the basis of substructure relationships between active compounds [3,4]. For substructure-based AC definition, the matched molecular pair (MMP) concept [5,6] has been proven to be very useful [3,4]. An MMP represents a pair of compounds that are only distinguished by a single chemical modification [5,6]. Following MMP terminology, the chemical modification (exchange of two substituents) is referred to as a transformation [5]. MMPs can be effectively generated algorithmically [5], which is an added plus for AC exploration, enabling efficient large-scale analysis. For substructure-based AC definition, MMPs with size-restricted transformations have been generated [7]. Size-restricted transformations represent typical R-group replacements in medicinal chemistry, leading to the introduction of MMP-cliffs [7]. Importantly, MMP-cliffs focus AC analysis on structurally analogous compound pairs and analog series. Series of structural analogs with significant potency variations preferentially form coordinated ACs, in other words, multiple and overlapping ACs [8,9]. Most ACs are formed in a coordinated manner, rather than by isolated pairs of compounds [8].

To further increase medicinal chemistry relevance and synthetic accessibility, MMPs can also be generated by bond fragmentation on the basis of retrosynthetic criteria such as RECAP rules [10]. Accordingly, structural modifications in such MMPs are synthesis-based. MMPs generated on the basis of retrosynthetic fragmentation

newlands press

10.4155/fmc-2018-0299 © 2019 Newlands Press ISSN 1756-8919

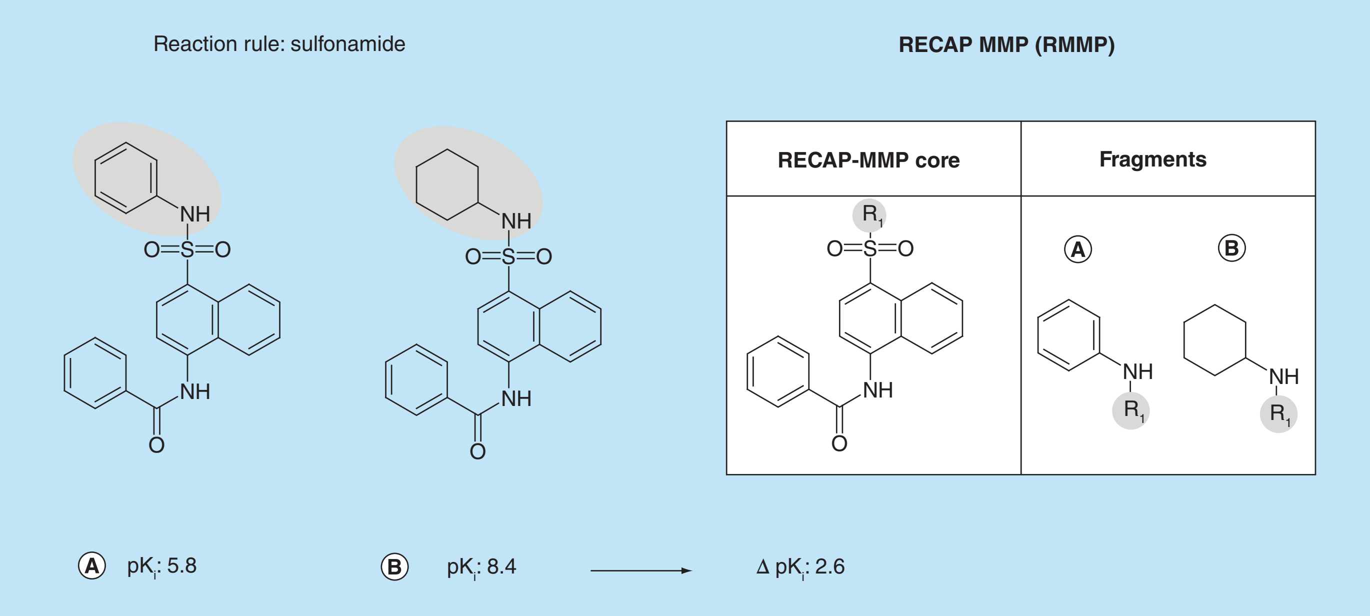


Figure 1. Retrosynthetic matched molecular pair-cliff. Shown is an exemplary RMMP-cliff formed by two C–C chemokine receptor type 8 antagonists (compounds **A** and **B**). The structural modification distinguishing the cliff partners is highlighted on a gray background, the reaction rule leading to bond fragmentation and RMMP formation is given and the logarithmic potency difference between the two compounds is reported.
RMMP: Retrosynthetic matched molecular pair.

were termed retrosynthetic (RECAP) MMPs (RMMPs) [11]. RMMPs can also be used to define ACs, resulting in RMMP-cliffs [12], in analogy to MMP-cliffs.

In addition to the similarity criterion, the potency difference criterion for AC formation must be clearly defined. Different from calculated similarity, potency values and resulting potency differences are experimentally determined properties. However, the comparison of different types of potency measurements must be avoided. For example, (assay-dependent) pIC_{50} and (assay-independent) pK_i values cannot be directly compared. If available, equilibrium constants are preferred potency measurements for ACs. Importantly, ACs have so far been defined on the basis of constant potency difference thresholds that were generally applied across target sets [2]. Specifically, on at least 100-fold difference in potency has often been used as a potency difference criterion for AC formation [2]. Our preferred general AC definition requires the formation of a transformation size-restricted MMP (or RMMP) with at least a 100-fold difference in potency between the paired structural analogs. Figure 1 shows an exemplary RMMP-cliff.

Applying the general MMP-cliff definition to target sets in ChEMBL [13], approximately 25% of all bioactive compounds, for which high-confidence activity data were available, were found to be involved in the formation of at least one AC. In addition, approximately 6% of all MMPs formed by compounds with high-confidence activity data represented ACs [14]. These findings provided insights into AC frequency when a general definition was applied across different targets sets.

Recently, we have gone a step further and analyzed compound potency distributions in target sets and their influence on AC formation [12]. Potency distributions of different variability were identified and assigned to three statistically distinct categories. Narrow potency distributions comprising category (CAT) 1 were unlikely to yield ACs, whereas more variable CAT 2 and CAT 3 distributions often yielded ACs, depending on intra-set structural relationships [12]. Taken together, the findings of our analysis suggested that potency difference criteria for ACs should be determined in a target set-dependent manner, as a complement or alternative to a generally applied potency difference criterion.

Herein, we present, for the first time, an assessment of ACs on the basis of target set-dependent potency difference thresholds, leading to a new set-dependent AC definition. Second-generation ACs further increase cliff information for target sets and extend the AC concept.

Materials & methods

Compounds & activity data

ChEMBL version 23 [13] was used as a source of compounds with well-defined (high-confidence) activity measurements applying the following selection criteria:

- Compounds with direct human target interactions (type 'D') and highest assay confidence (assay confidence score 9);
- Specified K_i values (equilibrium constants).

Application of these criteria yielded 71,967 unique compounds that were assigned to 904 target sets.

RMMP analysis

For RMMP generation, the following conditions were applied:

- Transformation size restrictions [7];
- Bond fragmentation following retrosynthetic rules (RECAP rules) [10,11].

A total of 354,094 target set-based RMMPs (243,110 unique RMMPs) were generated that involved 46,977 unique compounds from 574 target sets. Only target sets with at least 100 RMMPs were subjected to AC analysis (237 sets).

This RMMP threshold was applied to ensure that statistically meaningful potency distributions were obtained for RMMP-forming compounds. The 237 remaining sets yielded a total of 347,025 target-based RMMPs (238,795 unique RMMPs) involving 44,451 distinct RMMP-forming compounds.

Interquartile ranges of potency distributions

Following our previously established classification scheme [12], potency distributions of the 237 target sets were assigned to different categories on the basis of the interquartile range (IQR) in box plots capturing the distributions. The IQR represents the potency range between quartile 1 (Q1) and 3 (Q3) for approximately 50% of the compounds per set, as shown in Figure 2 (top). The following observations were made:

- CAT 1: IQR less than tenfold potency difference;
- CAT 2: IQR 10- to $<$100-fold potency difference;
- CAT 3: IQR $\geq$100-fold potency difference.

By definition, the IQR captured the potency range of approximately 50% of the compounds in each target set. CAT 1 potency distributions were typically narrow, as illustrated in Figure 2 (bottom), and hence, unlikely to yield ACs, except due to questionable outliers.

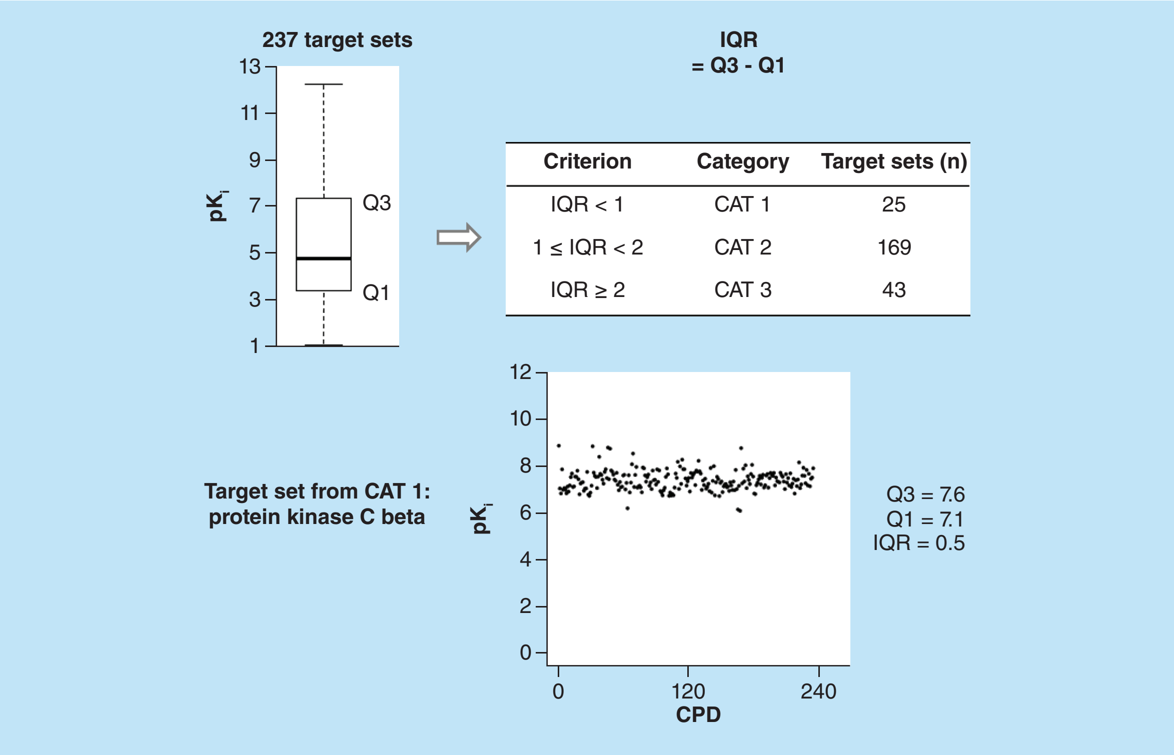
AC definition

Two alternative potency difference criteria were applied for defining RMMP-cliffs. The general (target set-independent) AC definition required a potency difference (Δ pK_i) of at least two orders of magnitude between RMMP partner compounds, as discussed above. For the newly introduced target set-dependent AC definition, qualifying potency differences between RMMP partners were set to the mean plus at least two standard deviations (σ) of the potency distribution among RMMPs of a given set, as shown in Figure 3.

As a structural similarity criterion for AC formation, we required the formation of a transformation size-restricted RMMP (instead of a conventional MMP). Thus, ACs identified herein were designated RMMP-cliffs.

Compound networks

An RMMP network was generated for each target set (nodes: compounds, edges: pairwise RMMPs) in which disjoint clusters were formed by individual analog series. RMMP networks were used to represent RMMP-cliffs by highlighting corresponding edges. All network representations were drawn with Cytoscape [15].



Criterion	Category	Target sets (n)
IQR < 1	CAT 1	25
1 ≤ IQR < 2	CAT 2	169
IQR ≥ 2	CAT 3	43

Figure 2. Potency distribution-dependent target set categories. The compound potency distributions of 237 target sets were monitored in a boxplot (top left) and the IQR, in other words, the difference between quartile 3 and 1, was determined. On the basis of the IQR, target sets were assigned to three different categories (CAT 1: IQR $<$ 1; CAT 2: 1 $\leq$ IQR $<$ 2; CAT 3: IQR $\geq$ 2). An exemplary potency distribution for a target set (protein kinase C beta inhibitors) belonging to CAT 1 is shown at the bottom.

CAT: Category; CPD: Compound; IQR: Interquartile range;

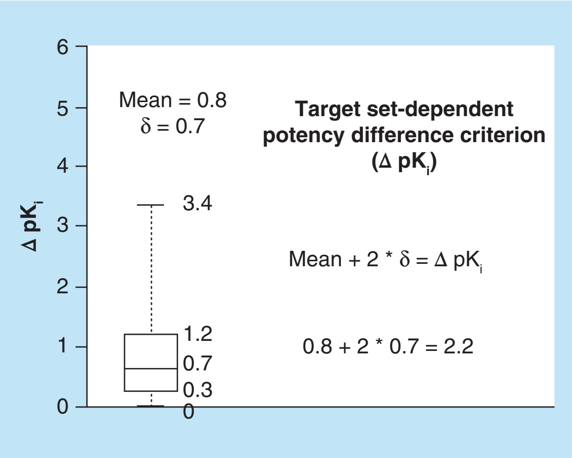


Figure 3. Target set-dependent potency difference criterion. For an exemplary target set (acetylcholinesterase inhibitors), potency differences between RMMP partners are monitored in a boxplot. In addition, the mean and standard deviation (σ) of the distribution were calculated. The target set-dependent potency difference threshold for activity cliff formation was calculated as the mean plus two σ (Δ pK_i = 2.2). RMMP: Retrosynthetic matched molecular pair.

Results & discussion

Formal criteria for AC definition & assessment

Defining and representing ACs in a consistent and generally applicable manner requires consideration of compound similarity criterion and potency difference criterion. Similarity can be evaluated using different compound representations and similarity functions [2,3]. We developed a preference for substructure-based similarity assessment [7], which ultimately resulted in the introduction of RMMP-cliffs [11,12], emphasizing medicinal chemistry relevance and synthetic accessibility. While AC research has so far preferentially focused on the similarity criterion and AC representation, less attention has been paid to the experimentally grounded potency difference criterion.

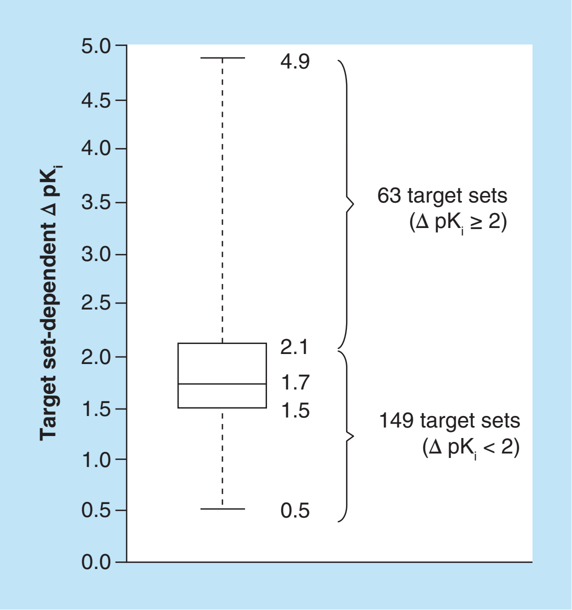


Figure 4. Distribution of target set-dependent potency difference thresholds. The global distribution of target set-dependent potency difference threshold values determined for 212 sets according to Figure 3 is shown in a boxplot. Values of the lower whisker, first quartile, median, third quartile and upper whisker are given. In addition, the number of target sets with potency difference thresholds of Δ pK$_i$ $\geq$ 2 and Δ pK$_i$ $<$ 2, respectively, is reported.

However, the use of high-confidence activity data has been emphasized to ensure that ACs convey reliable SAR information [3,4]. A generally applicable definition of ACs requires setting the magnitude of the potency difference threshold to a constant value that must be met or exceeded. This approach has dominated AC analysis over the years. By contrast, no target set-dependent AC definition has so far been introduced. Compound potency distributions strongly influence SARs and differences between distributions affect AC formation. Taking such differences into account requires the assessment of ACs at the level on individual target sets, which has motivated the introduction of a target set-dependent AC definition applying a constant similarity criterion and structural representation (Figure 1) and a variable potency difference criterion.

Potency distributions

Introducing target set-dependent potency difference threshold values for AC definition was expected to adapt ACs to target set characteristics and further increase the relevance of ACs for SAR exploration. Systematic analysis of compound potency distributions across target sets has revealed different levels of intra-set variability that were categorized on the basis of distribution statistics [12], as illustrated in Figure 2. CAT 1 sets with IQR values smaller than one have narrow potency distributions and are unlikely to yield ACs, except due to outliers, which should be considered with caution. By contrast, potency distributions of CAT 2 and 3 sets with larger IQR values are variable and likely to yield ACs, depending on intra-set structural relationships. Therefore, these sets have high priority for AC investigation. Whether or not target sets yield ACs depends on their SAR features and not all sets are expected to contain ACs. We determined that 25 of the 237 target sets that were preselected for our analysis had narrow potency distributions belonging to CAT 1. Accordingly, these sets were omitted and 212 CAT 2 and 3 sets remained for AC analysis.

target set-dependent potency differences

Next, we determined potency difference distributions for these 212 target sets by calculating potency differences for compound pairs forming RMMPs in each set, as illustrated in Figure 3. The median, mean and standard deviation of the potency difference distributions were determined.

In most instances, adding two standard deviations to the mean of the potency difference distributions resulted in values significantly exceeding the third quartile of the distributions, as shown in Figure 3. Therefore, the ‘mean plus two σ’ was set as a potency difference criterion to calculate target set-dependent threshold values for RMMP-cliff definition.

Figure 4 shows the resulting distribution of target set-dependent thresholds that included 63 sets with Δ pK$_i$ $\geq$ 2 and 149 sets with 0.5 $\leq$ Δ pK$_i$ $<$ 2. The median value of the Δ pK$_i$ distribution was 1.7 and thus similar to the potency difference threshold of 2.0 typically set for globally assessing ACs. The third quartile boundary of the

Table 1. Retrosynthetic matched molecular pair-cliff statistics.

Definition	General	target set-dependent
Target sets with RMMP-cliffs, n	195	212
RMMP-cliffs, n	11,773	16,096
RMMP-cliffs (mean per set), %	3.6	4.9
RMMP-cliffs (range per set), %	0.2–40.9	1.1–8.3
Coordinated RMMP-cliffs, %	92.8	92.7

RMMP-cliff statistics are compared for the general and targetset-dependent definition.
RMMP: Retrosynthetic matched molecular pair.

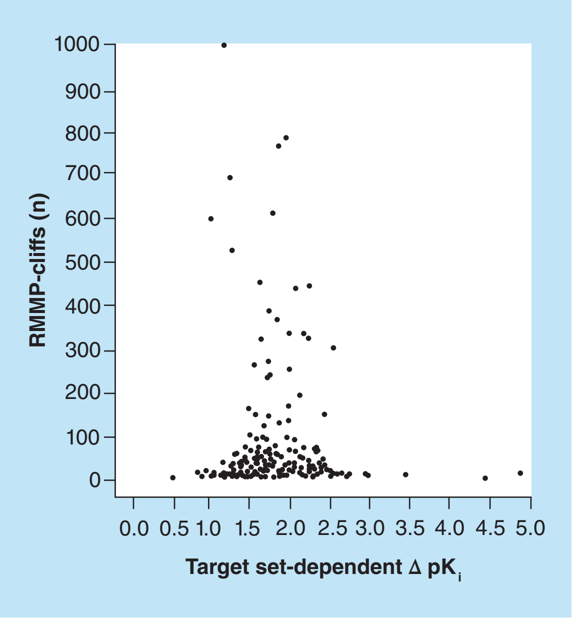


Figure 5. Distribution of RMMP-cliffs over set-dependent potency difference thresholds. For each target set, the number of RMMP-cliffs is plotted against the set-dependent potency difference threshold.
RMMP: Retrosynthetic matched molecular pair.

distribution was 2.1 (Figure 4). Thus, approximately 25% of the Δ pK_i values exceeded the potency difference threshold generally applied for AC analysis.

target set-dependent formation of ACs

RMMP-cliffs were defined in a target set-dependent manner by setting the potency difference criterion to the 'mean plus two σ' of the RMMP-based potency difference distribution. As reported in Table 1, a total of 16,096 RMMP-cliffs were obtained across all 212 target sets. Figure 5 shows the distribution of RMMP-cliffs over calculated set-dependent potency difference thresholds. The majority of target sets contained fewer than 100 RMMP-cliffs. However, in some target sets, much larger numbers of RMMP-cliffs were detected. Irrespective of the number of RMMP-cliffs per target set, the majority of target set-dependent potency difference thresholds fell into the Δ pK_i interval (1.5–2.5). Therefore, to further evaluate AC frequency as a function of target set-dependent potency difference thresholds, control calculations were carried out. In these calculations, a 'mean plus one σ' difference threshold criterion was applied, which resulted in more than 48,000 RMMP-cliffs across all target sets. This strong increase in ACs further supported the application of the statistically derived and more conservative 'mean plus two σ' criterion for target set-dependent threshold calculations.

Comparison of target set-dependent and generally defined ACs

Target set-dependent RMMP-cliffs were compared with those obtained on the basis of the general definition (Δ $pK_i \geq 2$). The results are summarized in Table 1. Generally defined RMMP-cliffs were detected in 195 of 212 target sets, involving a total of 7948 (12.6%) unique cliff-forming compounds. No cliffs were detected in 17 sets. All 212 CAT 2 and 3 target sets contained set-dependent cliffs that were generated by 11,167 (17.7%) unique cliff-forming compounds. Overall, more target set-dependent RMMP-cliffs (i.e., 16,096) than generally

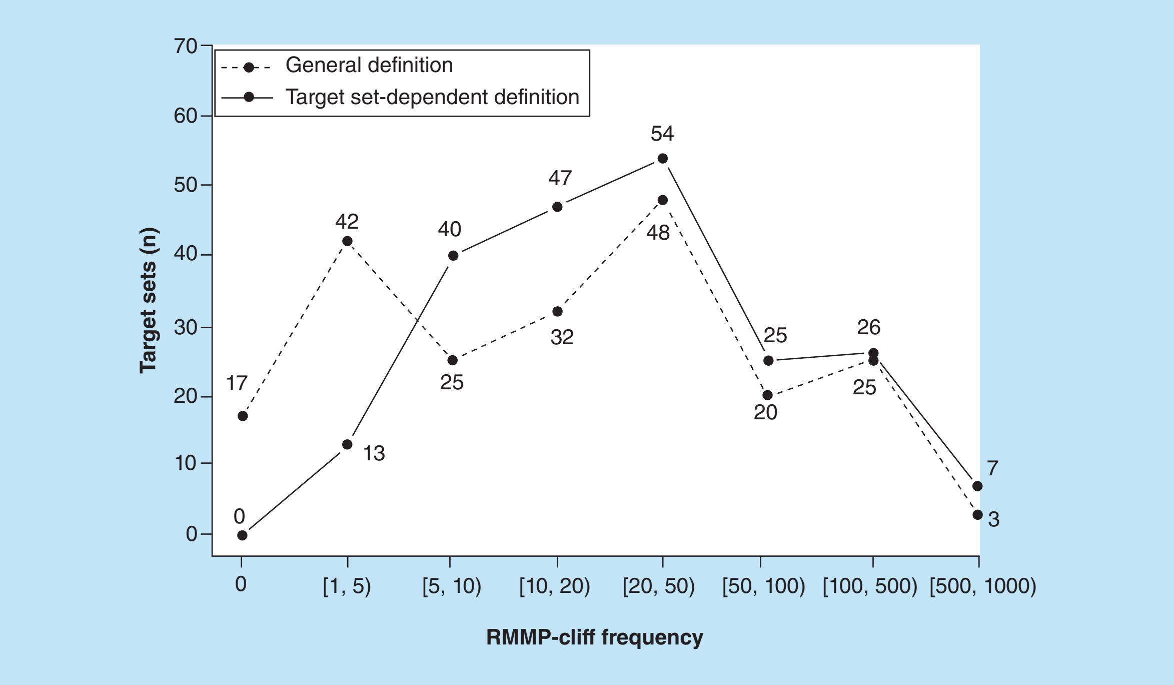


Figure 6. RMMP-cliff frequency for alternative definitions. The frequency of RMMP-cliffs applying the general definition (constant potency difference criterion of at least two orders of magnitude; dashed line) and the target set-dependent definition (solid line) is reported using different bins (x-axis). For each data point, the corresponding number of target sets falling into the frequency bin is given.
RMMP: Retrosynthetic matched molecular pair.

defined cliffs (11,773) were identified, with a nearly identical proportion of coordinated cliffs (92.7 vs. 92.8%). 61 target sets contained fewer set-dependent than generally defined RMMP-cliffs whereas 147 sets contained more set-dependent cliffs. In only four target sets, both definitions yielded the same number of RMMP-cliffs. Thus, the target set-dependent definition significantly modified RMMP-cliff populations across target sets. Figure 6 compares RMMP-cliff frequency on a target set basis and highlights definition-dependent variations in cliff frequency.

Applying the general definition, 3.6% of all RMMPs formed cliffs, with a range of 0.2–40.9% per set; applying the target set-dependent definition, 4.9% of all RMMPs represented cliffs, with a range of 1.1–8.3% per set (Table 1). Thus, as illustrated in Figure 7, the target set-dependent definition produced a much more balanced distribution of RMMP-cliffs across target sets than the general definition, both for sets containing small and large numbers of RMMPs. This was a direct consequence of applying the statistically derived target set-dependent potency difference criterion.

Figure 8A–C shows the potency distribution of compounds involved in the formation of RMMPs (top left) and the distribution of potency differences among RMMPs (bottom left) for three exemplary target sets. In addition, RMMP-cliff distributions resulting from the application of the general and target set-dependent AC definitions are compared. The target sets differed in the number of RMMP-forming compounds (Figure 8A: 62, B: 167 and C: 258 compounds) and in the corresponding potency ranges (Figure 8A: 4.7, B: 5.7 and C: 8.2). The number of RMMP-forming compounds did not correlate with the number of RMMPs (Figure 8A: 502, B: 534 and C: 605 RMMPs). In addition, potency differences captured by RMMPs were variable. For example, the RMMP-forming compounds reported in Figure 8C spanned a potency range of more than eight orders of magnitude. However, only six of 605 RMMPs exceeded a potency difference of more than two orders of magnitude, which resulted in the lower potency difference criterion for AC formation for the given target set. By contrast, RMMP-forming compounds from the target set in Figure 8A spanned a potency range of four orders of magnitude but RMMPs in this target set displayed a wider range of potency differences, with 59 RMMPs exceeding Δ p$K_i \geq 2$. This resulted in a larger potency difference criterion for AC formation for this target set. Furthermore, corresponding RMMP network representations are compared highlighting RMMP-cliffs resulting from the alternative AC definitions. Differences in AC formation indicate changes in SAR information content depending on the applied AC definition. In

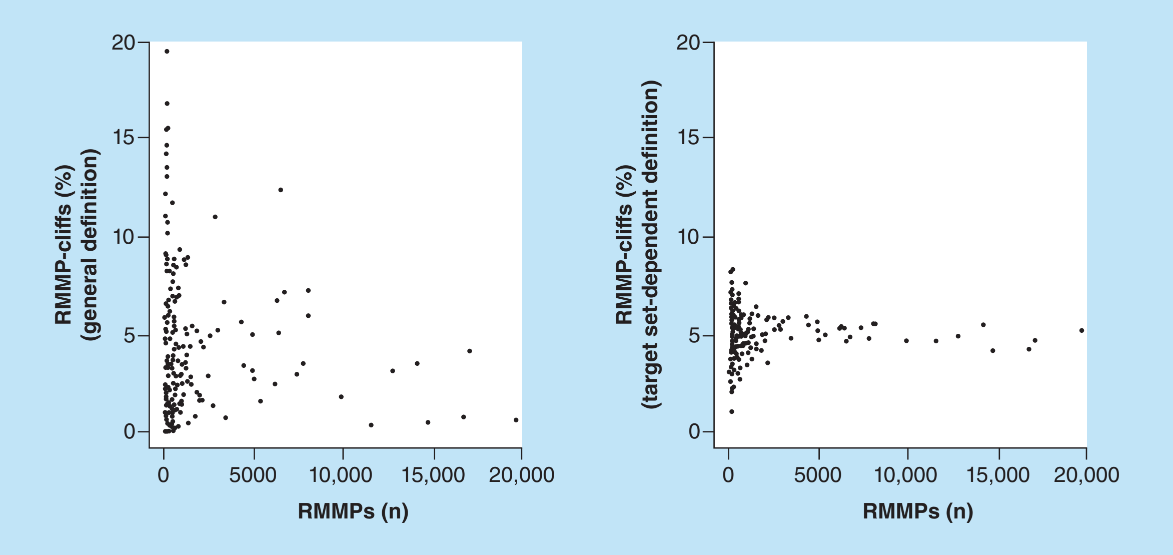
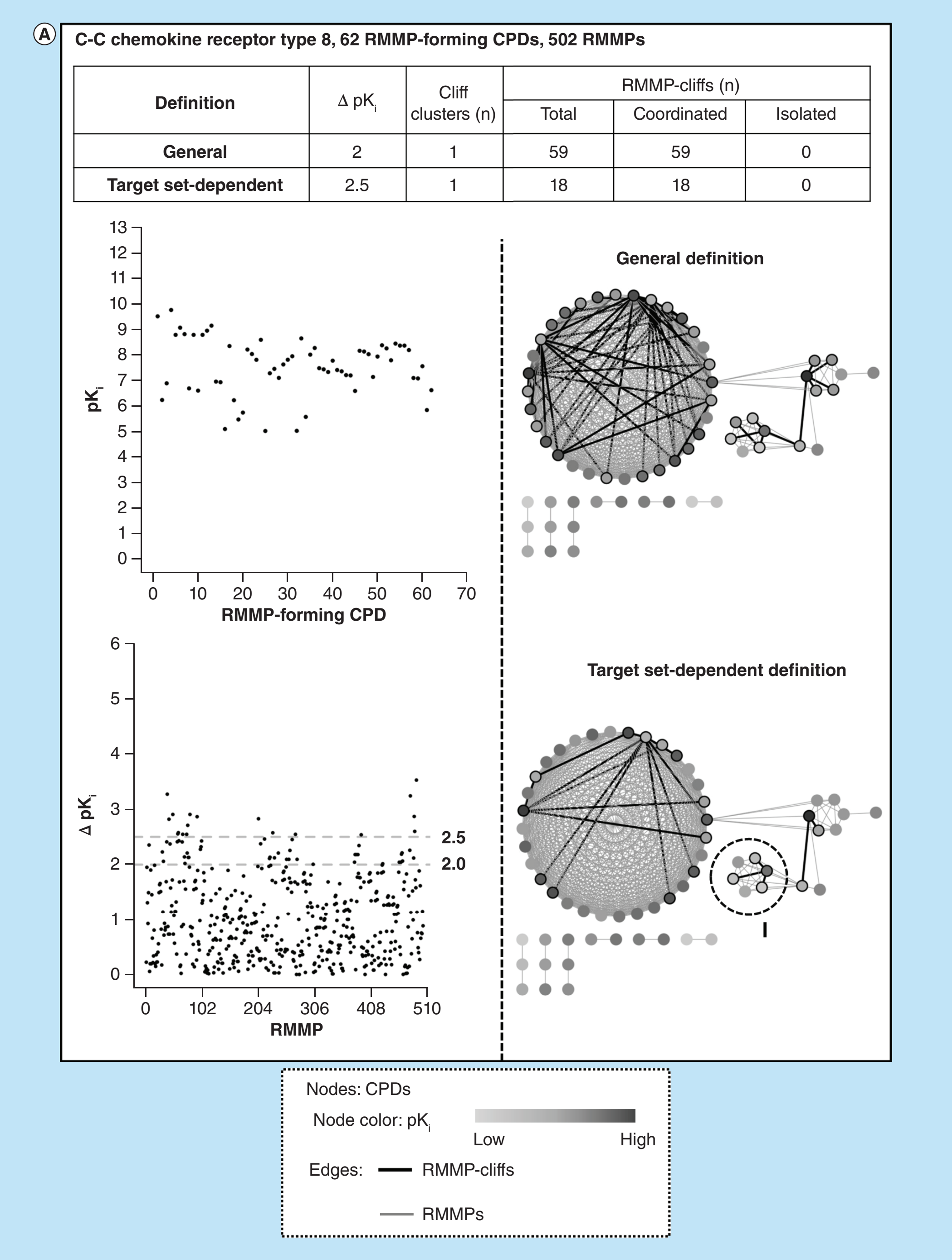


Figure 7. Proportion of RMMPs forming activity cliffs. Two scatter plots (general definition, left; target set-dependent definition, right) report the proportion of RMMPs forming RMMP-cliffs (RMMP-cliffs %) for each target set. For clarity, two outlier sets with very large percentages were omitted.
RMMP: Retrosynthetic matched molecular pair.

Figure 8A, a target set is depicted for which the set-dependent potency difference threshold (Δ pK_i = 2.5) exceeded the generally applied threshold (Δ pK_i = 2.0), which resulted in the formation of fewer set-dependent than generally defined RMMP-cliffs. Comparison of the RMMP networks shows that cliff formation in this structurally homogenous set was, in both instances, mostly confined to a large cluster of structural analogs with extensive RMMP relationships. In Figure 8B, a target set is shown for which the set-dependent potency difference threshold (Δ pK_i = 1.7) was smaller than the general threshold, resulting in more target set-dependent cliffs. Similar to the set in Figure 8A, RMMP-cliff formation in this structurally homogenous set was also mostly confined to a large compound cluster with extensive RMMP relationships. By contrast, in Figure 8C, a structurally heterogeneous target set is depicted yielding an RMMP network with diversified cluster structure. As in Figure 8C, the target set-dependent potency threshold (Δ pK_i = 1.4) was smaller in this case than the general threshold. Comparison of the resulting RMMP-cliff populations revealed that the increase in target set-dependent RMMP-cliffs led to cliff formation in several clusters representing different compound subsets. This observation was frequently made for structurally heterogeneous target sets with increasing numbers of set-dependent RMMP-cliffs. Hence, in these cases, cliff formation occurred in different structural contexts, thereby increasing the SAR information associated with set-dependent RMMP-cliffs compared with generally defined RMMP-cliffs; an important characteristic of set-dependent cliffs. Figure 8D shows exemplary set-dependent RMMP-cliffs for all three target sets together with their RMMP cluster locations.

Figure 9 shows examples of RMMP-cliffs that were formed if the general AC definition and the target set-dependent definition were applied. These RMMP-cliffs involved a single highly potent compound (on the left in Figure 9) and varying numbers of weakly potent analogs. When the general definition was applied, 10 RMMP-cliffs were obtained with similar potency differences and a variety of R-group replacements, giving rise to obvious redundancy in AC information. If the target set-dependent potency difference criterion Δ pK_i = 2.2 was applied only six of these 10 RMMP-cliffs were obtained (with weakly potent cliff compounds shown in the right column of Figure 9), which revealed essentially the same SAR information. Hence, in this case, the target set-dependent decrease in the number of RMMP-cliffs reduced redundancy and balanced AC information. In total, 484 generally defined and 323 target set-dependent RMMP-cliffs were obtained for this target set.

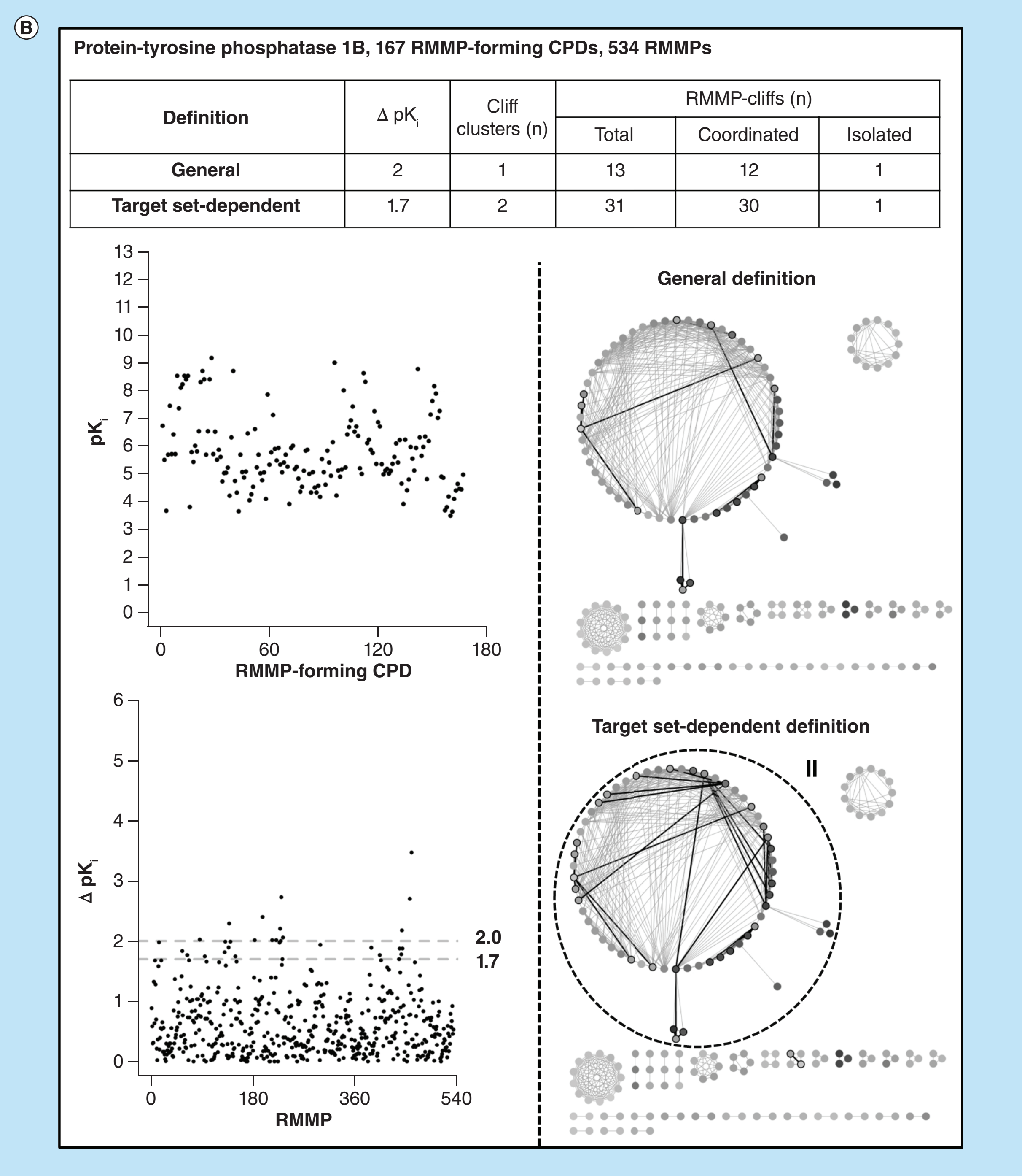
Figure 10 shows four exemplary RMMP-cliffs for another target set. Three of these RMMP-cliffs (except the one at the top in Figure 10) were only detected if the target set-dependent potency difference criterion Δ pK_i = 1.3 was applied. These RMMP-cliffs further increased the associated SAR information content because, as can be seen, they were formed in different structural contexts. In total, this target set yielded only two generally defined and 28



Definition	Δ pK$_i$	Cliff clusters (n)	RMMP-cliffs (n)		
			Total	Coordinated	Isolated
General	2	1	59	59	0
Target set-dependent	2.5	1	18	18	0

Figure 8. Comparison of definition-dependent activity cliff formation. For individual target sets (target name given at the top), the potency value distribution for RMMP-forming compounds (CPDs) is shown (left, upper plot) and potency differences between RMMP partners are reported (left, lower plot). Dashed gray lines indicate potency difference threshold values for the general AC definition (Δ pK$_i$ ≥ 2, constant) and the target set-dependent definition (variable, e.g. Δ pK$_i$ ≥ 2.5 in 8**A**). In addition, RMMP networks are shown after applying the general (right, upper representation) and target set-dependent definition (right, lower representation). Network nodes are colored by potency using a continuous spectrum (shown at the bottom) from light gray (lowest compound potency in the target set) to black (highest potency). RMMP-cliffs are highlighted using thick edges connecting nodes. Individual clusters from which exemplary RMMP-cliffs are shown in 8**D** are encircled and numbered. Furthermore, RMMP-cliff statistics are reported (top). **(A)** Decrease in RMMP-cliffs. Details are provided for a target set for which the application of the target set-dependent definition resulted in a decrease in the number of RMMP-cliffs relative to the general definition. **(B)** and **(C)** Increase in RMMP-cliffs. Details are given for target sets with an increase in the number of RMMP-cliffs for the target set-dependent relative to the general definition. **(D)** Exemplary RMMP-cliffs. Shown are RMMP-cliffs applying the target set-dependent definition from clusters encircled in **(A)–(C)**.

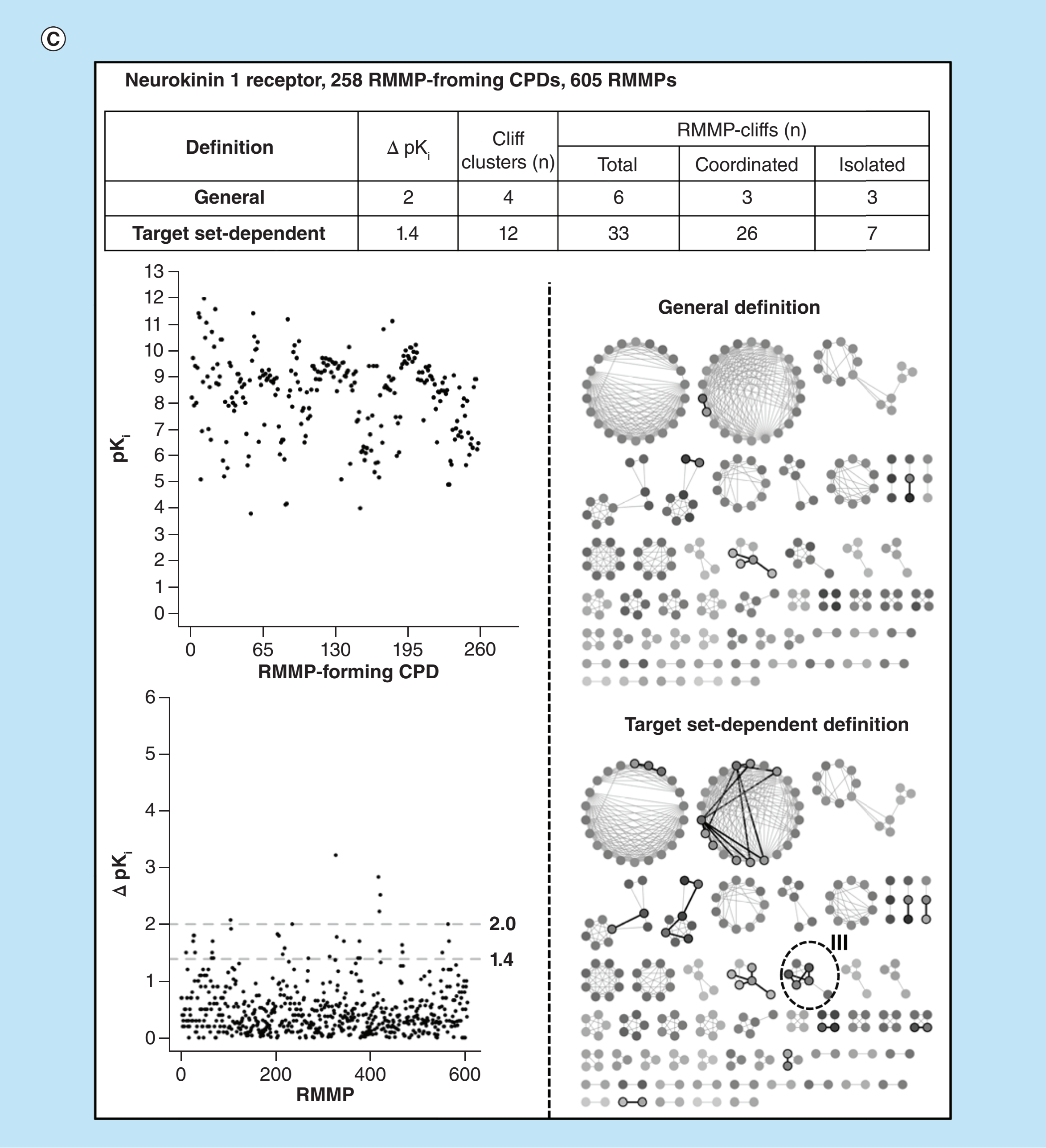
AC: Activity cliff; CPD: Compound; RMMP: Retrosynthetic matched molecular pair.



Definition	Δ pK_i	Cliff clusters (n)	RMMP-cliffs (n)		
			Total	Coordinated	Isolated
General	2	1	13	12	1
Target set-dependent	1.7	2	31	30	1

Figure 8. Comparison of definition-dependent activity cliff formation (cont.). For individual target sets (target name given at the top), the potency value distribution for RMMP-forming compounds (CPDs) is shown (left, upper plot) and potency differences between RMMP partners are reported (left, lower plot). Dashed gray lines indicate potency difference threshold values for the general AC definition (Δ p$K_i \geq 2$, constant) and the target set-dependent definition (variable, e.g. Δ p$K_i \geq 2.5$ in 8**A**). In addition, RMMP networks are shown after applying the general (right, upper representation) and target set-dependent definition (right, lower representation). Network nodes are colored by potency using a continuous spectrum (shown at the bottom) from light gray (lowest compound potency in the target set) to black (highest potency). RMMP-cliffs are highlighted using thick edges connecting nodes. Individual clusters from which exemplary RMMP-cliffs are shown in 8**D** are encircled and numbered. Furthermore, RMMP-cliff statistics are reported (top). **(A)** Decrease in RMMP-cliffs. Details are provided for a target set for which the application of the target set-dependent definition resulted in a decrease in the number of RMMP-cliffs relative to the general definition. **(B)** and **(C)** Increase in RMMP-cliffs. Details are given for target sets with an increase in the number of RMMP-cliffs for the target set-dependent relative to the general definition. **(D)** Exemplary RMMP-cliffs. Shown are RMMP-cliffs applying the target set-dependent definition from clusters encircled in **(A)–(C)**.

AC: Activity cliff; CPD: Compound; RMMP: Retrosynthetic matched molecular pair.



Neurokinin 1 receptor, 258 RMMP-froming CPDs, 605 RMMPs

Definition	Δ pK_i	Cliff clusters (n)	RMMP-cliffs (n)		
			Total	Coordinated	Isolated
General	2	4	6	3	3
Target set-dependent	1.4	12	33	26	7

Figure 8. Comparison of definition-dependent activity cliff formation (cont.). For individual target sets (target name given at the top), the potency value distribution for RMMP-forming compounds (CPDs) is shown (left, upper plot) and potency differences between RMMP partners are reported (left, lower plot). Dashed gray lines indicate potency difference threshold values for the general AC definition (Δ $pK_i \geq 2$, constant) and the target set-dependent definition (variable, e.g. Δ $pK_i \geq 2.5$ in 8**A**). In addition, RMMP networks are shown after applying the general (right, upper representation) and target set-dependent definition (right, lower representation). Network nodes are colored by potency using a continuous spectrum (shown at the bottom) from light gray (lowest compound potency in the target set) to black (highest potency). RMMP-cliffs are highlighted using thick edges connecting nodes. Individual clusters from which exemplary RMMP-cliffs are shown in 8**D** are encircled and numbered. Furthermore, RMMP-cliff statistics are reported (top). **(A)** Decrease in RMMP-cliffs. Details are provided for a target set for which the application of the target set-dependent definition resulted in a decrease in the number of RMMP-cliffs relative to the general definition. **(B)** and **(C)** Increase in RMMP-cliffs. Details are given for target sets with an increase in the number of RMMP-cliffs for the target set-dependent relative to the general definition. **(D)** Exemplary RMMP-cliffs. Shown are RMMP-cliffs applying the target set-dependent definition from clusters encircled in **(A)–(C)**.

AC: Activity cliff; CPD: Compound; RMMP: Retrosynthetic matched molecular pair.

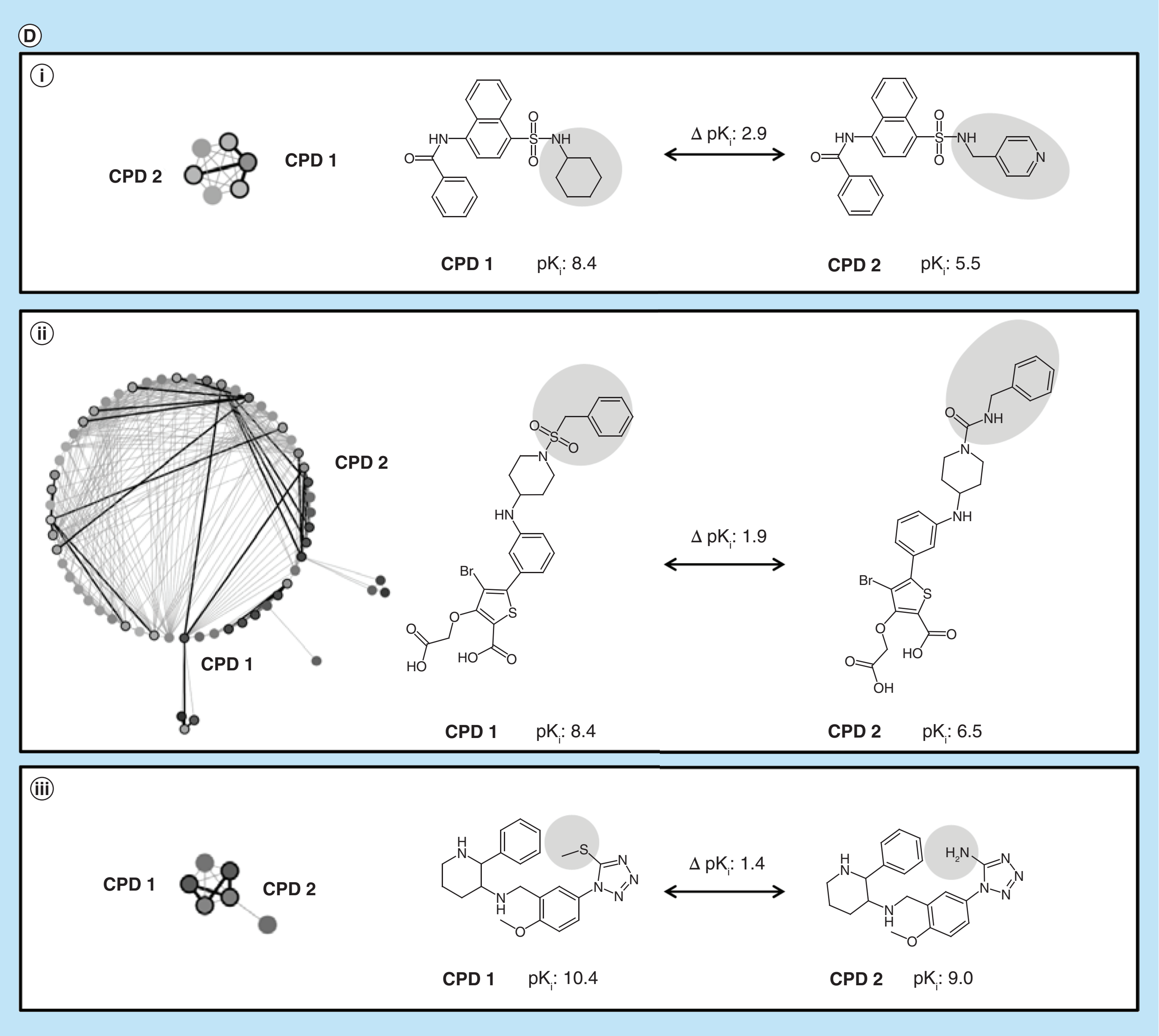


Figure 8. Comparison of definition-dependent activity cliff formation (cont.). For individual target sets (target name given at the top), the potency value distribution for RMMP-forming compounds (CPDs) is shown (left, upper plot) and potency differences between RMMP partners are reported (left, lower plot). Dashed gray lines indicate potency difference threshold values for the general AC definition (Δ pK$_i$ ≥ 2, constant) and the target set-dependent definition (variable, e.g. Δ pK$_i$ ≥ 2.5 in 8**A**). In addition, RMMP networks are shown after applying the general (right, upper representation) and target set-dependent definition (right, lower representation). Network nodes are colored by potency using a continuous spectrum (shown at the bottom) from light gray (lowest compound potency in the target set) to black (highest potency). RMMP-cliffs are highlighted using thick edges connecting nodes. Individual clusters from which exemplary RMMP-cliffs are shown in 8**D** are encircled and numbered. Furthermore, RMMP-cliff statistics are reported (top). **(A)** Decrease in RMMP-cliffs. Details are provided for a target set for which the application of the target set-dependent definition resulted in a decrease in the number of RMMP-cliffs relative to the general definition. **(B)** and **(C)** Increase in RMMP-cliffs. Details are given for target sets with an increase in the number of RMMP-cliffs for the target set-dependent relative to the general definition. **(D)** Exemplary RMMP-cliffs. Shown are RMMP-cliffs applying the target set-dependent definition from clusters encircled in **(A)–(C)**.
AC: Activity cliff; CPD: Compound; RMMP: Retrosynthetic matched molecular pair.

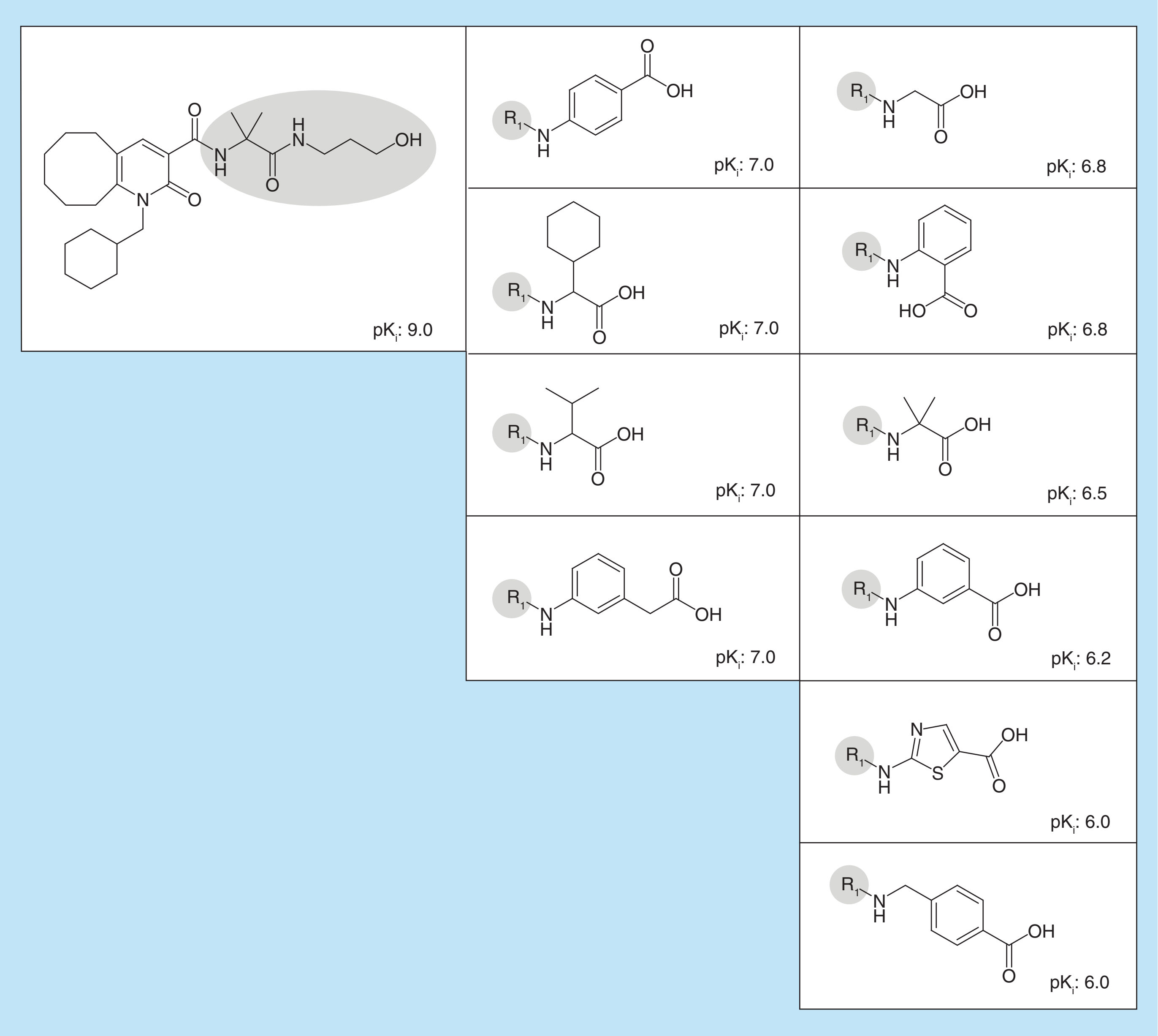


Figure 9. Balancing activity cliff redundancy. Shown are retrosynthetic combinatorial analysis. Procedure matched molecular pairs-cliffs from the cannabinoid CB2 receptor target set. Increasing the potency difference criterion for activity cliff formation from the generally applied Δ pK$_i$ ≥ 2 to the target set-dependent Δ pK$_i$ ≥ 2.2 led to a reduction in the number of highly similar activity cliffs.

target set-dependent RMMP-cliffs. Hence, in this case, given the target set-dependent potency distribution, the majority of SAR-informative compound pairs would have not been identified if the general AC definition had been applied.

Taken together, the examples in Figure 9 and Figure 10 further illustrate the utility and SAR relevance of second-generation ACs introduced herein.

Future perspective

In this work, we have explored ACs on the basis of target set-dependent potency value distributions and introduced a novel AC concept considering target set-dependent potency difference criteria. This extension of generally defined ACs is of high relevance for SAR analysis because potency value distributions vary significantly across target sets. Narrow potency distributions in target sets are unlikely to yield ACs. The presence of variable potency distributions is

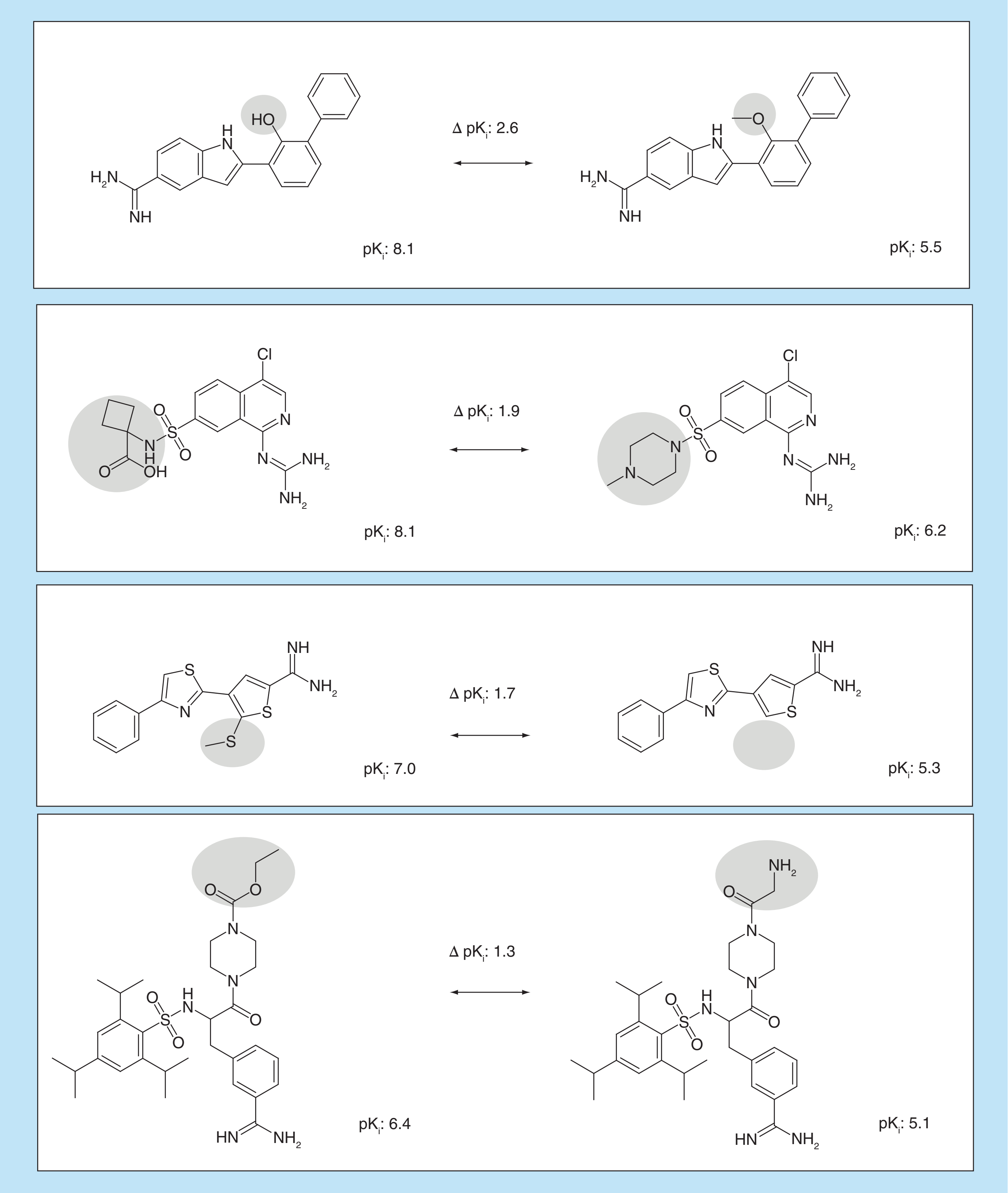


Figure 10. Increase in activity cliff-associated structure–activity relationship information. Shown are retrosynthetic matched molecular pair-cliffs from the urokinase-type plasminogen activator target set. Decreasing the potency difference criterion for activity cliff from the generally applied Δ pK$_i$ = 2.0 to a target set-dependent Δ pK$_i$ = 1.3 increased the number of RMMP-cliffs and associated structure–activity relationship information. This was the case because activity cliffs with different structural contexts were identified.

a necessary but insufficient condition for AC formation, which is also strongly influenced by structural relationships between active compounds. For example, if all analogs in a given structurally unique subset of a target set have high potency and all analogs in another unique subset have low potency, no ACs are formed, despite strong potency variations within the set. However, taking differences in potency distributions into account adapts AC analysis to target set-specific features, and hence, increases the SAR relevance of the analysis. To these ends, the first target set-dependent AC definition has been introduced herein. On the basis of statistical analysis, a target set-dependent potency difference criterion was derived and applied to calculate target set-dependent potency difference thresholds for AC formation. For qualifying target sets with statistically significant potency variations, the potency difference criterion for set-dependent AC formation was set to the mean plus at least two standard deviations of the potency difference distribution of RMMPs. For the majority of target sets, a larger number of set-dependent than generally defined RMMP-cliffs was obtained, and the distribution of RMMP-cliffs was more balanced across target sets. For structurally diverse target sets, a relative increase in set-dependent RMMP-cliffs also resulted in AC formation across different compound subsets. Thus, ACs were formed in a different structural context provided by compounds from different subsets, which led to an increase in AC-associated SAR information. Thus, target set-dependent definition and assessment of ACs further supports SAR analysis and extends the utility of AC information for medicinal chemistry applications.

ACs have been and continue to be of high interest in medicinal chemistry. It is anticipated that the introduction of second-generation ACs will lead to a more extensive exploration of ACs in target sets of interest, especially as new target sets evolve. Adding AC information from public compound sources to medicinal chemistry projects will provide new opportunities for compound optimization. Making this information available in the practice of medicinal chemistry will depend on the involvement of computational scientists who are capable of working with rapidly increasing amounts of compound and activity data. Thus, investigators trained in data science and chemistry might be highly sought after in the future to operate at the interface between computational and medicinal chemistry. Regardless, for systematically exploring ACs and associated SAR information, a methodological foundation has been laid and the key criteria have been thoroughly investigated. Hence, the consideration of second-generation ACs is expected to provide many future opportunities for practical applications in medicinal chemistry.

Summary points

Formal criteria for activity cliff definition & assessment

- The definition of activity cliffs (AC) requires the specification of similarity and potency difference criteria.
- General definitions have been applied so far on the basis of given molecular representations and similarity measures and constant potency difference threshold values.

Variable potency distributions

- Determining target set-dependent potency difference thresholds was expected to further increase the relevance of ACs for structure–activity relationship (SAR) exploration.
- Potency value distributions displayed high variability across target sets.

target set-dependent potency differences

- Analysis of potency difference distributions for pairs of structurally analogous compounds in target sets enabled the derivation of a target set-dependent potency difference criterion for AC formation.

target set-dependent formation of activity cliffs

- A target set-dependent definition of ACs was introduced by applying a constant similarity criterion and variable potency difference thresholds.

Comparison of target set-dependent & generally defined activity cliffs

- The target set-dependent definition yielded more ACs than the general definition and a more balanced distribution of ACs across target sets.
- An increase in the number of set-dependent ACs compared with general ACs was often accompanied by AC formation in different compound subsets, thereby increasing AC-associated SAR information content.

Future perspective

- Target set-dependent ACs account for different potency value distributions and improve the utility of ACs for SAR analysis.
- The target set-dependent definition of ACs further extends the AC concept in medicinal chemistry.

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

References

Papers of special note have been highlighted as: • of interest

1. Maggiora GM. On outliers and activity cliffs – why QSAR often disappoints. *J. Chem. Inf. Model.* 46(4), 1535–1535 (2006).

• **Original definition of activity cliffs.**

2. Stumpfe D, Bajorath J. Exploring activity cliffs in medicinal chemistry. *J. Med. Chem.* 55(7), 2932–2942 (2012).

• **Review of the activity cliff concept in medicinal chemistry.**

3. Stumpfe D, Hu Y, Dimova D, Bajorath J. Recent progress in understanding activity cliffs and their utility in medicinal chemistry. *J. Med. Chem.* 57(1), 18–28 (2014).

• **Review of recent extensions of the activity cliff concept.**

4. Bajorath J. Evolution of the activity cliff concept for structure–activity relationship analysis and drug discovery. *Future Med. Chem.* 6(14), 1545–1549 (2014).

5. Hussain J, Rea C. Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *J. Chem. Inf. Model.* 50(3), 339–348 (2010).

6. Griffen E, Leach AG, Robb GR, Warner DJ. Matched molecular pairs as a medicinal chemistry tool. *J. Med. Chem.* 54(22), 7739–7750 (2011).

7. 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.* 52(5), 1138–1145 (2012).

8. Stumpfe D, Dimova D, Bajorath J. Composition and topology of activity cliff clusters formed by bioactive compounds. *J. Chem. Inf. Model.* 54(2), 451–561 (2014).

9. Dimova D, Stumpfe D, Bajorath J. Method for the evaluation of structure–activity relationship information associated with coordinated activity cliffs. *J. Med. Chem.* 57(15), 6553–6563 (2014).

10. Lewell XQ, Judd DB, Watson SP, Hann MM. RECAP-retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* 38(3), 511–522 (1998).

11. de la Vega de León A, Bajorath J. Matched molecular pairs derived by retrosynthetic fragmentation. *Med. Chem. Commun.* 5(1), 64–67 (2014).

12. Hu H, Stumpfe D, Bajorath J. Rationalizing the formation of activity cliffs in different compound data sets. *ACS Omega* 3(7), 7736–7744 (2018).

13. Gaulton A, Bellis LJ, Bento AP *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40(database issue), D1100–D1107 (2012).

• **ChEMBL is the major repository of compounds and activity data from medicinal chemistry and major data source for activity cliff (AC) analysis.**

14. Stumpfe D, Bajorath J. Monitoring global growth of activity cliff information over time and assessing activity cliff frequencies and distributions. *Future Med. Chem.* 7(12), 1565–1579 (2015).

15. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27(3), 431–432 (2010).

Summary

Since potency distributions across activity classes are generally not comparable, a constantly applied potency difference criterion might not always be suitable for defining ACs. In this study, a formula for computing activity class-dependent potency difference thresholds was introduced by considering the mean plus two standard deviations of potency differences of RMMPs as a threshold for AC formation. Using RMMP as the structural similarity criterion, activity class-dependent potency difference thresholds frequently ranged from 1 to 2.5 orders of magnitude with a median value of 1.7 (ΔpK_i). The comparison of activity class-dependent and generally defined potency difference criteria indicated that the activity class-dependent AC definition generally yielded higher cliff numbers (16,096 vs. 11,773 RMMP-cliffs) with higher cliff target coverage (212 vs. 195 targets) and a balanced cliff percentage of (1.1-8.3% vs. 0.2-40.9%). Moreover, RMMP network analysis indicated that activity class-dependent potency difference criterion balanced AC formation in different AC clusters, thus enabling AC formation within diverse structural contexts.

In the following analysis, we systematically explored AC formation in analog series using the newly introduced activity class-dependent potency difference criteria.

Chapter 9

Introducing a New Category of Activity Cliffs with Chemical Modifications at Multiple Sites and Rationalizing Contributions of Individual Substitutions

Introduction

For graph-based molecular similarity evaluation, matched molecular pairs (MMPs), also termed single-site analog pairs, have been intensively studied. Herein, we analyzed AC characteristics by considering ACs from analog series (AS). ASs were computationally identified using the recently introduced compound-core relationship (CCR) algorithm. Since the CCR methodology permits at most five non-hydrogen substitution sites for each compound, paired analogs from the same AS might be distinguished by one or multiple substitution sites. Multi-site ACs were obtained by extracting analog pairs from ASs and their formation was rationalized.

Reprinted with permission from “Stumpfe, D.; Hu, H.; Bajorath, J. Introducing a New Category of Activity Cliffs with Chemical Modifcations at Multiple Sites and Rationalizing Contributions of Individual Substitutions. *Bioorganic & Medicinal Chemistry* **2019**, *27*, 3605-3612”. *Copyright 2019 Elsevier*

Bioorganic & Medicinal Chemistry 27 (2019) 3605–3612

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

Introducing a new category of activity cliffs with chemical modifications at multiple sites and rationalizing contributions of individual substitutions

Dagmar Stumpfe[1], Huabin Hu[1], Jürgen Bajorath*

Department of Life Science Informatics, b-it, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Endenicher Allee 19c, D-53115 Bonn, Germany

ARTICLE INFO

Keywords:
Activity cliffs
Structure-activity relationships
Analog series
Analog pairs
Potency differences
Multiple substitutions
Multi-site activity cliffs
Differential substitution effects

ABSTRACT

Activity cliffs (ACs) are formed by structurally similar active compounds with large potency differences. In medicinal chemistry, ACs are of high interest because they reveal structure-activity relationship (SAR) information and SAR determinants. Herein, we introduce a new type of ACs that consist of analog pairs with different substitutions at multiple sites (multi-site ACs; msACs). A systematic search for msACs across different classes of bioactive compounds identified more than 4000 of such ACs, most of which had substitutions at two sites (dual-site ACs; dsACs). A hierarchical analog data structure was designed to analyze contributions of individual substitutions to AC formation. Single substitutions were frequently found to determine potency differences captured by dsACs. Hence, in such cases, there was redundancy of AC information. In instances where both substitutions made significant contributions to dsACs, additive, synergistic, and compensatory effects were observed. Taken together, the results of our analysis revealed the prevalence of single-site ACs (ssACs) in analog series, followed by dsACs, which reveal different ways in which paired substitutions contribute to the formation of ACs and modulate SARs.

1. Introduction

In medicinal chemistry and chemical informatics, activity cliffs (ACs) are generally defined as structurally similar compounds that share the same activity but have large differences in potency.[1,2] Thus, ACs reveal small chemical modifications having a profound effect on biological activity. Accordingly, ACs are important sources of structure-activity relationship (SAR) information.[2,3]

For identifying and studying ACs, molecular similarity relevant for AC formation and the magnitude of potency differences that qualify compound pairs as ACs must unambiguously defined.[2,3] This has been done in different ways, reflecting an evolution of the AC concept over the years.[4] Originally, ACs were defined by quantifying compound similarity through calculation of the Tanimoto coefficient (Tc) using molecular fingerprints as descriptors, a key technique in chemical informatics.[2,4] For AC definition, Tc calculations have some limitations. First, a numerical threshold value for similarity must be subjectively defined. Second, Tc-based compound similarity is not related to chemical reactions and often difficult to reconcile from a medicinal chemistry perspective.

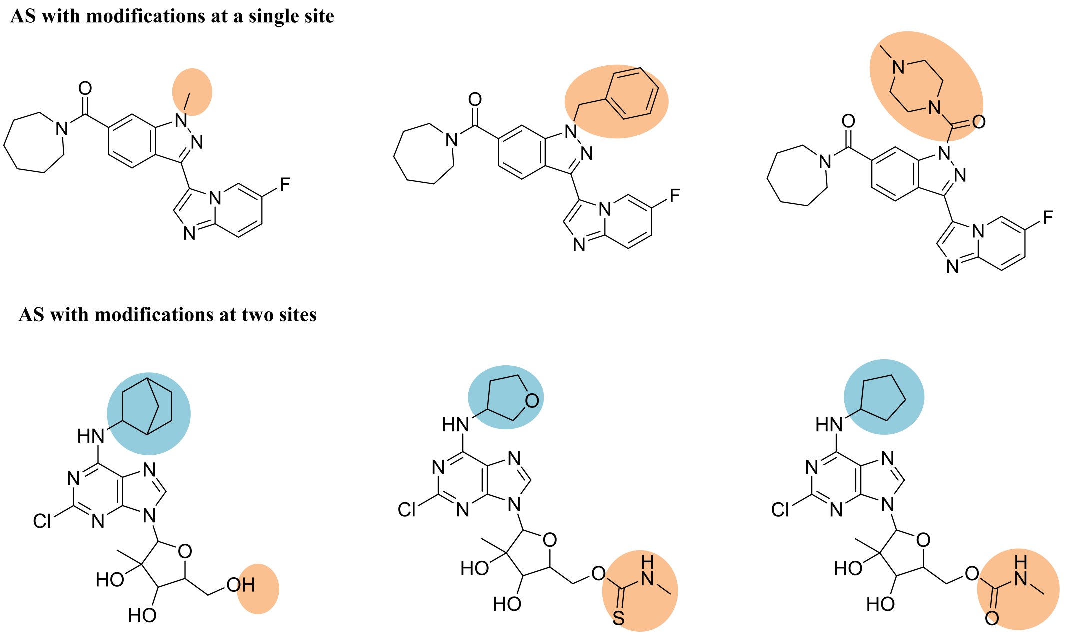
Instead of threshold-based calculations, substructure-based similarity criteria have also been applied such as, for example, the presence of shared molecular scaffolds[4] or the formation of matched molecular pairs (MMPs),[4,5] i.e. pairs of compounds that are only distinguished by a chemical modification at a single site.[5,6] Applying meaningful size restrictions for chemical modifications,[7] MMP calculations have been used to identify pairs of compounds for AC exploration, leading to the introduction of MMP-cliffs.[7] For AC analysis, random deletion of bonds in compounds to generate MMPs[6] has also been replaced by fragmentation according to retrosynthetic rules,[8,9] which has further increased the chemical interpretability of MMP-cliffs in light of reaction information.[9] While Tc-based ACs might include multiple structural changes, depending on a given pair of similar compounds, a characteristic feature of MMP-cliffs is that they only carry modifications as a single site.[7,9]

In addition to evaluating molecular similarity relationships, a potency difference threshold must be determined to identify ACs.[2,3] The consistent application of a similarity criterion and potency difference threshold is essential for unambiguously analyzing ACs,[3,4] especially when searching for ACs across different data sets.[10] A potency difference of at least two orders of magnitude (100-fold) has often been applied as an AC criterion,[4,10] given that potency differences of this

* Corresponding author.
E-mail address: bajorath@bit.uni-bonn.de (J. Bajorath).

[1] The contributions of these authors should be considered equal.

https://doi.org/10.1016/j.bmc.2019.06.045
Received 21 May 2019; Received in revised form 24 June 2019; Accepted 27 June 2019
Available online 28 June 2019


AS with modifications at a single site

AS with modifications at two sites

Fig. 1. Analog series with different number of substitution sites. Exemplary analogs from two different series of adenosine A2a receptor ligands are shown that differ in the number of substitution sites. Analogs from the series at the top are distinguished by modifications at a single substitution site (highlighted in orange), while analogs from the series at the bottom have modifications at two sites (highlighted in blue and orange, respectively).

magnitude are considered significant in medicinal chemistry.

The general application of a constant potency difference threshold does not take into consideration that potency value distributions often significantly differ across compound activity classes (also called target sets), which affects AC formation.[11] Therefore, we have recently introduced the derivation of target set-dependent potency difference thresholds, which takes set-dependent potency variations into account and focuses AC analysis on most significant potency differences.[12] This has led to the introduction of second-generation MMP-cliffs with target set-dependent potency difference thresholds.[12]

Herein, we report a further extension of the AC concept by introducing analog pair-based ACs with different substitutions at multiple sites. Computational identification of analog series[13,14] in target sets is followed by systematically enumerating all possible analog pairs. Single-site ACs (ssACs) and multi-site ACs (msACs) are then defined on the basis of target set-dependent potency differences.

In systematic search calculations, more than 4000 msACs were identified in more than 140 qualifying target sets. More than 90% of the msACs had different substitutions at two sites and were termed dual-site ACs (dsACs). For these ACs, a hierarchical analog data structure was designed to investigate contributions of individual substitutions to AC formation. ssACs were frequently found to represent dsACs. For dsACs where both substitutions significantly contributed to AC formation, additive, synergistic, and compensatory effects of substitutions were detected, thus making these ACs informative test cases for SAR exploration.

2. Materials, methods, and analysis concepts

2.1. Compounds and activity data

Bioactive compounds were extracted from ChEMBL version 24.1.[15] Only compounds with reported direct interactions (target relationship type: "D") with human targets at the highest confidence level (assay confidence score: 9), numerically defined equilibrium constants (K_i values), and exact measurements ("=") were selected. Equilibrium constants were reported as pK_i values. To ensure accuracy of AC assignments it is essential to limit the use of potency values to exact measurements. Given these criteria, a total of 73,965 unique compounds with activity against 915 different targets were obtained. Compounds with reported activity against each target were combined into an individual target set.

2.2. Target set-dependent potency distributions

For each of the 915 target sets, compound potency (pK_i) distributions were analyzed in boxplots. The interquartile range (IQR) of each potency distribution was determined, i.e., the value range covering the intermediate ~50% of the compounds.[11] Target sets were only further considered if their IQR was at least 1 (one order of magnitude). This criterion was applied because target sets with smaller IQR rarely yield ACs.[11] Accordingly, 525 target sets qualified for further analysis.

2.3. Determination of analog series

For each of the 525 target sets, a systematic search for analog series (ASs) was carried out applying a recently introduced computational methodology.[14] In each case, compounds were systematically decomposed by bond fragmentation according to retrosynthetic combinatorial analysis procedure (RECAP) rules,[8] permitting a maximum of five substitution sites per compound. Each fragmentation step produced a compound core and substituent fragment. A core was generally required to have at least twice the size (number of non-hydrogen atoms) of the fragment or combined multiple fragments. For each compound, all possible core-fragment combinations with single to at most five substitution sites were retained and substituents at each site were replaced by a hydrogen atom, thereby establishing a compound-core relationship (CCR) for each site.[14] Subsequently, all compounds sharing the same core were organized into an individual AS. Accordingly, compounds belonging to the same AS were distinguished by modifications at a single and/or multiple substitution sites. Fig. 1 shows exemplary analogs from two different ASs of adenosine A2a receptor ligands with a single and with two substitution sites. For 410 of the 525 qualifying target sets, a total of 15,087 target set-based ASs were identified.

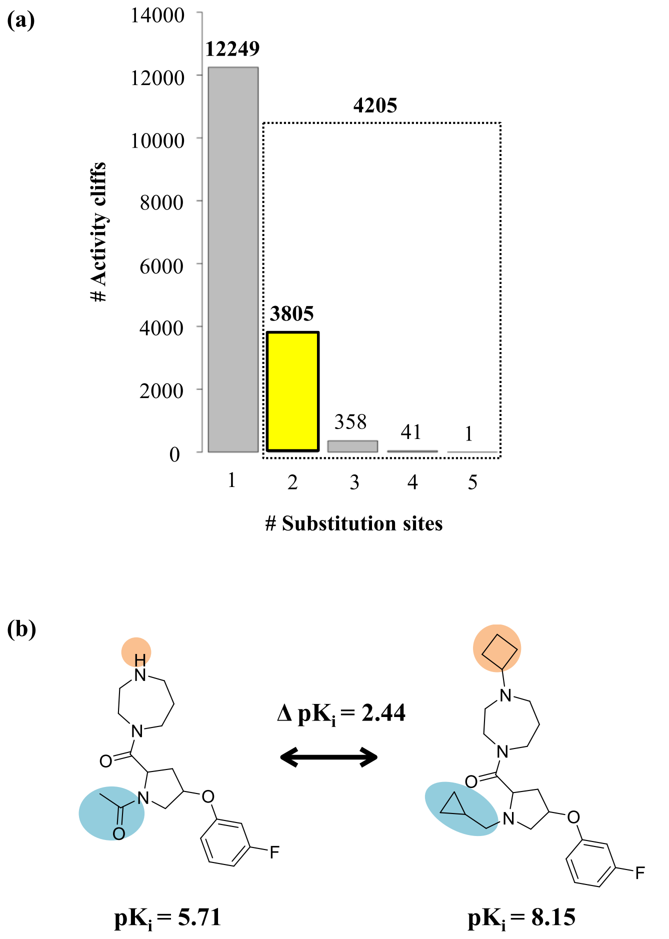


Fig. 2. Activity cliffs formed by analogs with single or multiple substitution sites. (a) The histogram reports the number of ACs formed by analogs distinguished at 1–5 substitution sites. The bar for dual-site ACs (dsACs), which are prevalent among multi-site ACs, is highlighted in yellow. (b) An exemplary dsAC formed by histamine H3 receptor ligands is shown.

2.4. Enumeration of analog pairs

By definition, all compounds belonging to one AS form pairwise analog relationships. For each target set, all analog pairs were extracted from its ASs. These analog pairs provided the basis for the identification of ACs, as described below. For our analysis, only target sets were further considered if they contained at least 100 analog pairs for AC analysis. This criterion reduced the number of target sets to 209, yielding a total of 14,065 target-based ASs comprising 39,540 unique compounds. These compounds formed a total of 334,306 analog pairs that were used to identify ACs.

2.5. Activity cliff definition and identification

For defining and evaluating ACs, a similarity criterion and potency difference criterion must be applied. In our analysis, two compounds were considered similar if they belonged to the same AS and hence formed an analog pair. Accordingly, the presence of a common structural core was required as a substructure-based similarity criterion for AC formation. An analog pair was then classified as an AC if the potency difference between the analogs was greater than or equal to the target set-dependent potency difference threshold.[12] In contrast to a generally applied potency difference value for AC formation, target set-dependent potency difference threshold takes the specific potency value distribution of compounds having the same activity into account, which can significantly differ between target sets.[11] A target set-dependent threshold for AC formation was determined as the mean value of the potency differences of all analog pairs per set plus two standard deviations (sigma).[12] For each target set, all ACs were then determined. As reported below, ssACs were found to dominate the distribution of ACs and dsACs the distribution of msACs.

2.6. Single-site analogs

For dsACs, a systematic search for *single-site analogs* was carried out that contained one of the two substitutions of the dsAC. Effects of structural modifications in a given dsACs were separately considered if two single-site analogs were identified for the cliff. Single-site analogs were required to originate from the same target set as the corresponding dsACs and have high-confidence activity data. Search calculations for single-site analogs were carried out with the aid of the OpenEye chemistry toolkit.[16]

2.7. Activity cliff redundancy

If an individual substitution was identified on the basis of single-site analogs that fully accounted for the potency difference captured by a dsAC, the dsAC was considered *redundant* because the SAR information was already provided by an ssAC. By contrast, if single substitutions were not detected that fully accounted for dsAC formation, and both substitutions contributed to the cliff, the dsAC was classified as *confirmed*. To avoid boundary effects in assigning redundant dsACs, a 10% potency difference deviation (margin) was permitted. Thus, if a potency difference matched by a single-site analog was detected within −10% of the observed value or above, the corresponding dsAC was classified as redundant.

2.8. Differential contributions of individual and combined substitutions

For confirmed dsACs, the sum of the Δ pK_i values for substitutions converting the weakly potent cliff partner into the single-site analogs was compared to the Δ pK_i of the dsAC. Three different effects were distinguished for combined single-site modifications:

(i) *Additive effect*: the Δ pK_i sum of single-site substitutions was comparable (± 10%) to the Δ pK_i of the dsAC.
(ii) *Synergistic effect*: the Δ pK_i sum was lower than the Δ pK_i (−10%) of the dsAC.
(iii) *Compensatory effect*: the Δ pK_i sum was greater than the Δ pK_i (+10%) of the dsAC.

Table 1
Activity cliffs with varying number of substitution sites and single-site analogs.

Substitution sites (n)		Type of ACs	# Single-site analog(s)	# ACs	# Redundant dsACs	# Confirmed dsACs
Single	n = 1	ssAC	-	12,249	-	-
Multiple	*n = 2*	*dsAC*	*0*	*1012*	–	–
			1	*2496*	*764*	–
			2	*297*	*141*	*156*
	n ≥ 3	msAC	n. d.	400	–	–

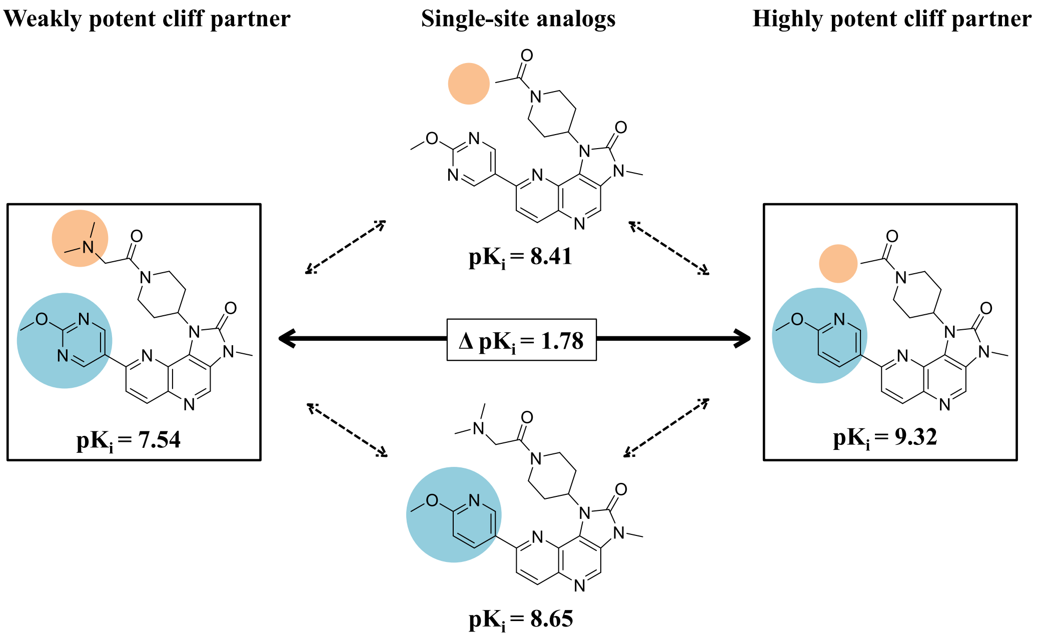


Fig. 3. Dual-site activity cliffs and single-site analogs. For an exemplary dsAC, two single-site analogs are shown that contain the individual substitutions found in the dsAC. Modifications at the two substitution sites are highlighted in blue and orange, respectively. The analogs are PI3-kinase p110-alpha subunit inhibitors. For each compound, the pK$_i$ value is reported. In this case, the target set-dependent potency difference (Δ pK$_i$) threshold for AC formation is 1.61. The solid arrow indicates the formation of the dsAC and dashed arrows indicate contributions of the individual substitutions. Analyzing individual contributions to multi-site ACs requires the availability of corresponding single-site analogs with potency measurements.

Thus, synergy increased the potency gain of combined individual substitutions, whereas compensation reduced the potency gain due to unfavorable combinatorial effects.

3. Results and discussion

3.1. Study concept

The study was designed to investigate a new category of ACs that are formed by pairs of analogs (substructure-based similarity criterion), meet target set-dependent potency difference thresholds, and contain multiple substitution sites. So-defined ACs were investigated for the first time. Therefore, ASs with single and/or multiple substitution sites were systematically extracted from target sets and set-dependent potency difference thresholds for ACs were derived. Then, a systematic search for ssACs and msACs was carried out across qualifying target sets. Subsequently, a search was conducted for single-site analogs of dsACs to determine contributions of individual substitutions to AC formation and differentiate effects of combined substitutions. Major aims of the study included the identification of newly defined msACs, exploration of relationships between ssACs and msACs, and elucidation of potency effects associated with individual and combined substitutions in ACs.

3.2. Distribution of activity cliffs with varying number of substitution sites

The final selection of 209 target sets yielded a total of 16,454 ACs with single or multiple substitution sites. These ACs included 12,249 ssACs and 4205 msACs, as reported in Fig. 2a and Table 1. Thus, the distribution of ACs across different target sets was dominated by ssACs, which was not expected. Yet, more than 25% of the detected ACs were msACs. Most of the msACs (3805 cliffs; 90.5%) were dsACs and only 400 msACs with three or more substitution sites were detected (Table 1). Therefore, our subsequent analysis focused on dsACs. An exemplary dsAC is shown in Fig. 2b.

Reported are the numbers of activity cliffs (# ACs) with varying substitution sites (n) including ssACs, dsACs, and msACs. Here dsACs (in *italics*) are the focal point. For dsACs, the number of ACs with 0, 1, or 2 available single-site analogs is reported. For dsACs with 1 or 2 single-site analogs, the number of redundant dsACs (# redundant dsACs) is given. In addition, for dsACs with two single-site analogs, the number of confirmed dsACs (# confirmed dsACs) is reported. msACs with three or more substitution sites were not further analyzed because their number was small compared to dsACs. "n.d." means not determined.

3.3. Dual-site activity cliffs and single-site analogs

In contrast to ssACs where the potency difference is directly attributable to a single substitution, contributions of individual substitutions to dsACs must be further analyzed. Therefore, a systematic search was carried out for single-site analogs of dsACs across the 136 qualifying target sets, which identified a total of 3090 single-site analogs.

Fig. 3 shows an exemplary dsAC for which two single-site analogs were identified. For the AC compounds and single-site analogs, pK$_i$ values are reported. The data structure depicted in Fig. 3 was used to investigate the contributions of individual substitutions to dsAC formation. Comparing potency differences between weakly potent dsAC compounds, single-site analogs, and highly potent dsAC compounds made it possible to search for individual substitutions that were responsible for the potency difference captured by a dsAC.

For 297 of the 3805 dsACs, both single-site analogs were identified. In addition, for 2496 other dsACs, one single-site analog was found. For the remaining 1012 dsACs, no single-site analog was detected.

3.4. Redundant and confirmed dual-site activity cliffs

For the 297 dsACs with two single-site analogs, the contributions of individual substitutions to AC formation were analyzed. As shown in Fig. 4, two different cases were distinguished in assigning redundant

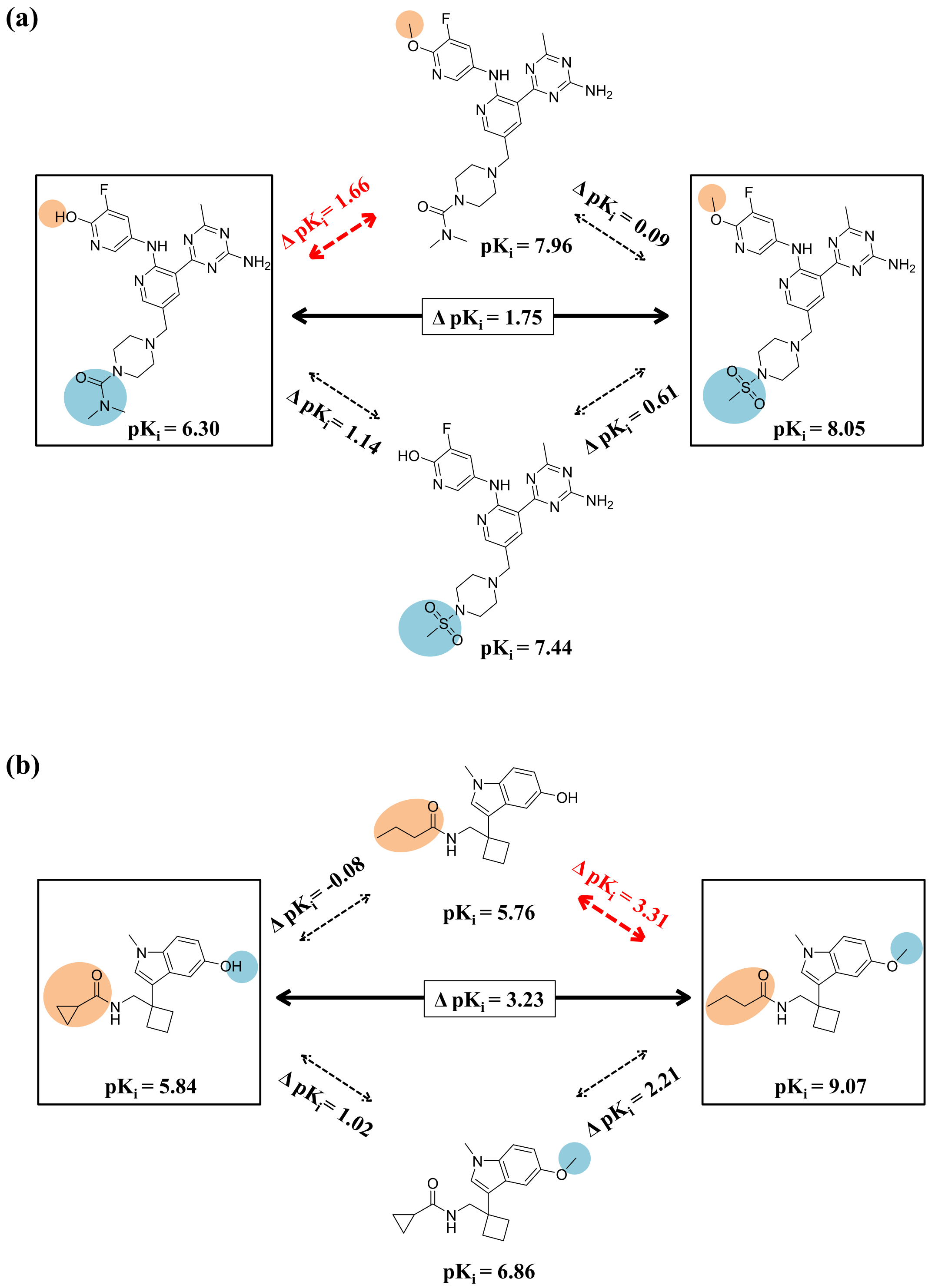
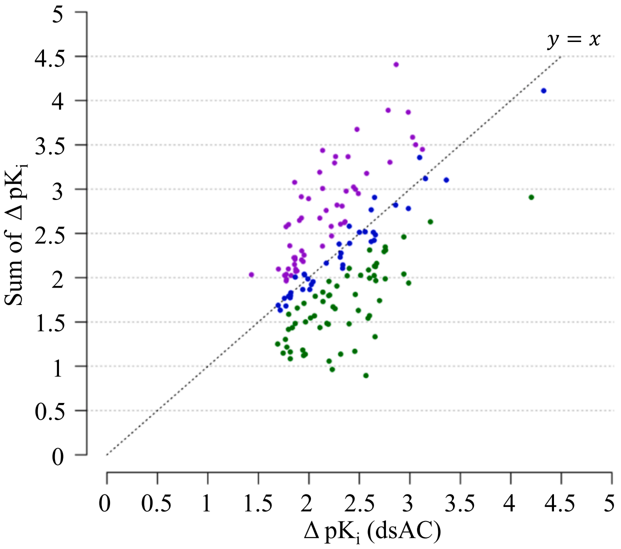


Fig. 4. Redundant dual-site activity cliffs. Two dsACs are shown that can be rationalized on the basis of ssACs (Δ pK$_i$ highlighted in red) formed between (a) the weakly potent cliff partner and a single-site analog or (b) a single-site analog and the highly potent cliff partner. The presentation is according to Fig. 3. The compounds in (a) originate from the PI3-kinase p110-alpha subunit target set (with a target set-dependent AC potency difference threshold of 1.61 pK$_i$ units) and the compounds in (b) from the melatonin receptor 1B target set (with a target set-dependent AC potency difference threshold of 2.59).

dsACs. First, a structural modification transforming a weakly potent cliff partner into a single-site analog might account for dsAC-based potency difference (Fig. 4a). Second, a structural modification of a single-site analog producing the highly potent cliff partner might also account for the potency difference (Fig. 4b). Both cases were considered to be equivalent.



🟢	Synergistic effect	62
🔵	Additive effect	38
🟣	Compensatory effect	56

Fig. 5. Varying effects of single-site substitutions in confirmed dual-site activity cliffs. For the 156 confirmed dsACs, the sum of the Δ pK_i values for converting the weakly potent cliff partner into the two single-site analogs is plotted against the Δ pK_i value of the dsAC, revealing additive (blue), synergistic (green), and compensatory (purple) effects. Each dot represents a dsAC.

We determined that 141 dsACs were covered by a corresponding ssAC; 41 of these ssACs were formed by a weakly potent dsAC compound and a single-site analog and 51 between a single-site analog and the highly-potent cliff partner. In addition, 49 dsACs were represented by multiple ssACs, involving both cliff partners and single-site analogs.

Furthermore, 764 of the 2496 dsACs for which only one single-site analog was available were found to be covered by an ssACs involving this single-site analog. For the remaining dsACs, it remained uncertain whether or not they might be represented by an ssAC because the second single-site analog was not available.

Taken together, the results of single-site analog analysis revealed a significant degree of redundancy for dsACs. Of the 297 dsACs for which both single-site analogs were available, 156 dsACs were confirmed, but nearly 50% were redundant (Table 1). Hence, considering the large number of ssACs compared to msACs that were initially identified and the large proportion of redundant dsACs (Table 1), AC formation was overall dominated by single substitutions.

In Fig. 4a, the potency difference of the dsAC (Δ pK_i 1.75) was already obtained (within a 10% margin) by replacing the hydrogen atom of the hydroxyl group of the weakly potent cliff partner by a methyl group (Δ pK_i 1.66). In Fig. 4b, a single-site analog formed an ssAC with the highly potent cliff partner, with a Δ pK_i of 3.31 (compared to 3.23 for the dsAC). Again, the replacement of the hydrogen atom of the hydroxyl group in the single-site analog by a methyl group produced the AC, but only if the tricyclic ring was already substituted. By contrast, separately considered individual substitutions of the weakly potent cliff partner did in this case not cover the potency difference of the dsAC. Thus, the structural context of the single-site analog was required for the methyl replacement to generate an ssAC covering the dsACs.

3.5. Potency effects of substitutions in confirmed dual-site activity cliffs

For the 156 confirmed dsACs, both substitutions generating single-site analogs were essential for the potency difference captured by the dsAC. For each confirmed dsAC, the sum of the potency differences of the individual substitutions was compared to the Δ pK_i of the dsACs. As shown in Fig. 5, the 156 confirmed dsACs were found to include 38 cliffs with additive, 62 with synergistic, and 56 with compensatory effects. Thus, although additivity of potency effects was observed, for more than 75% of the confirmed dsACs, the substitutions yielded synergistic or compensatory effects. Thus, dsACs cannot be confidently predicted by addition of potency changes from corresponding single-site analogs. Fig. 6 shows examples for additive (**6a**), synergistic (**6b**), and compensatory (**6c**) substitutions in confirmed dsACs, illustrating that dsACs and corresponding single-site analogs provide instructive examples for SAR exploration.

3.6. Substitution patterns in dual-site activity cliffs and different potency effects

Substitution combinations contained in all dsACs were systematically analyzed. The set of 3805 dsACs contained 3264 dsACs (85.8%) with unique combinations of substitutions. Hence, there was a high degree of diversity among dsAC substitution combinations and there were no substitution patterns that occurred with high frequency in dsACs. The most frequent combination was a pair of hydrogen atom vs. methoxy group (H/methoxy) replacements that was found in 54 unique dsACs, followed by a H/methoxy plus H/methyl combination in 26 unique dsACs and a pair of H/methyl substitutions in 15 unique dsACs. These substitutions were the smallest and hence most generic substitutions detected in dsACs. Overall, there were no preferred substitution combinations leading to dsAC formation.

For the subset of 156 confirmed dsACs, it was possible to relate substitution patterns to different potency effects. Corresponding to the observation made for all dsACs, the 156 confirmed dsACs included 131 dsACs (84.0%) with unique combinations. Only two to at most five confirmed dsACs were detected that shared the same combination. Interestingly, most of dsACs with a shared combination had different potency effects. For example, in four dsACs containing a pair of H/methoxy replacements, the individual substitutions had compensatory, synergistic, or additive effects, as described above. The same observations were made for four other dsACs containing a pair of H/methyl replacements. Thus, potency effects of the same substitution combinations in dsACs strictly depended on the target the cliff compounds were active against.

4. Conclusions

Herein, we have introduced a new category of ACs with multiple substitution sites based upon systematically identified pairs of analogs from many different series. For defining these ACs, target set-dependent potency difference criteria were applied. A systematic search for analog pair-based ACs war carried out, which identified more than 12,000 ssACs and 4000 msACs. Hence, ssACs were much more frequent than msACs. Furthermore, more than 90% of the newly identified msACs were dsACs. These dsACs were found in nearly 65% of all qualifying target sets. To analyze contributions of individual substitutions to dsAC formation, a data structure was devised combining dsACs with single-site analogs representing individual substitutions. A systematic search identified single-site analogs for about 2800 dsACs, thus enabling further detailed analysis. For 297 dsACs, both single-site analogs were detected. Surprisingly, nearly half of these dsACs were found to be represented by ssACs, revealing redundancy among dsACs. Taken together, our findings revealed a clear dominance of ssACs over msACs, which was not anticipated. For confirmed dsACs, different effects of combined substitutions were identified and synergistic or compensatory effects outnumbered additive contributions. Although ssACs are more frequent than msACs, newly identified dsACs further extend the current spectrum of ACs. However, our study also shows that significant potency variations among analogs from medicinal chemistry preferentially result from single substitutions, which has interesting implications for compound optimization efforts, suggesting thorough exploration of individual substitution sites before multiple site are considered simultaneously.

Acknowledgements

The authors would like to thank Martin Vogt for helpful discussion and the OpenEye Free Academic Licensing Program for providing a free academic license for the chemistry toolkit. H.H. is supported by the China Scholarship Council (CSC).

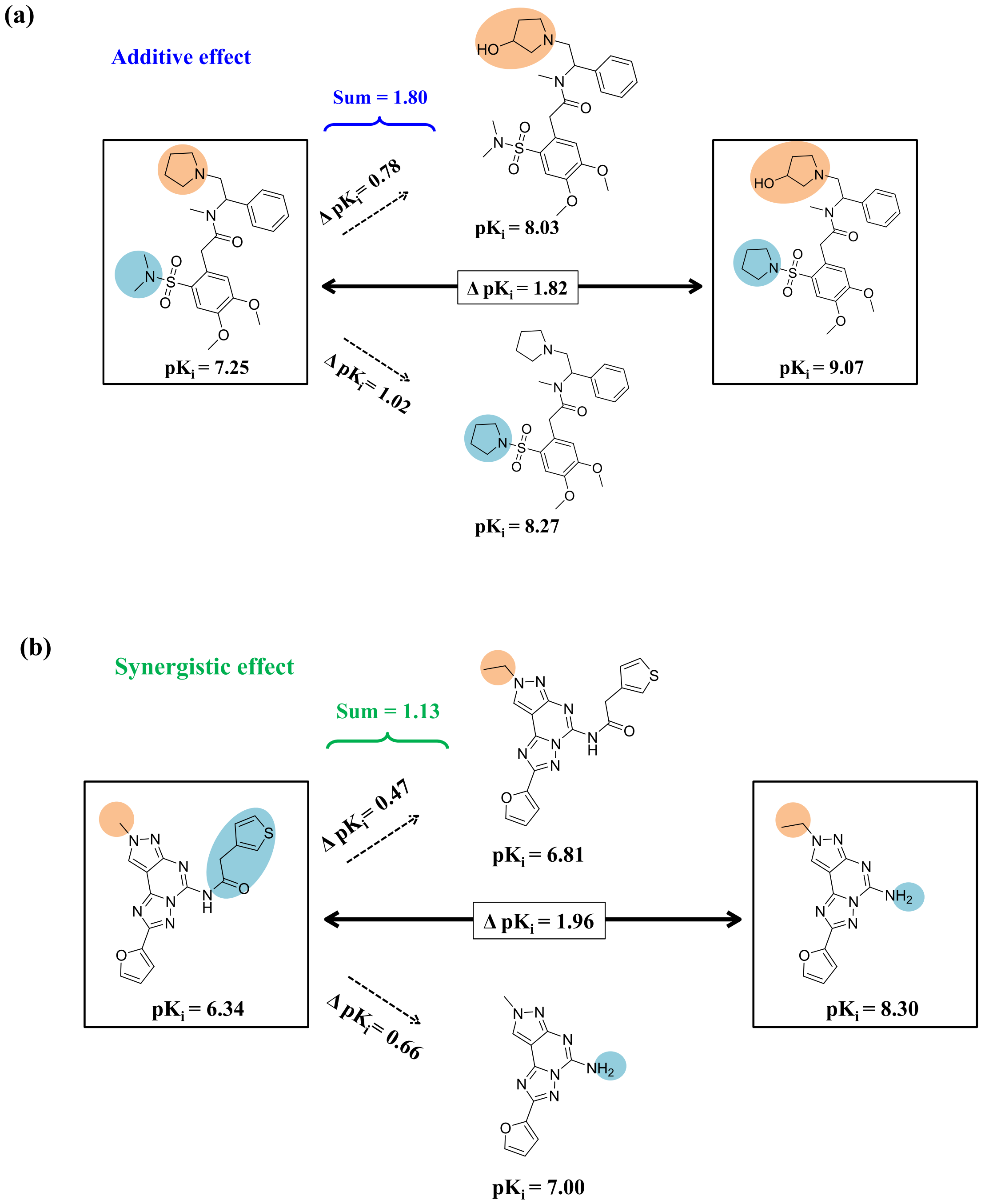


Fig. 6. Additive, synergistic, and compensatory effects in dual-site activity cliffs. Shown are exemplary dsACs with (a) additive, (b) synergistic, and (c) compensatory effects of single-site substitutions. The representation is according to Fig. 3. The dsACs in (a), (b), and (c) are taken from the delta opioid receptor (target set-dependent AC potency difference threshold of 1.70), adenosine A1 receptor (1.90) and progesterone receptor (2.07) target set, respectively.

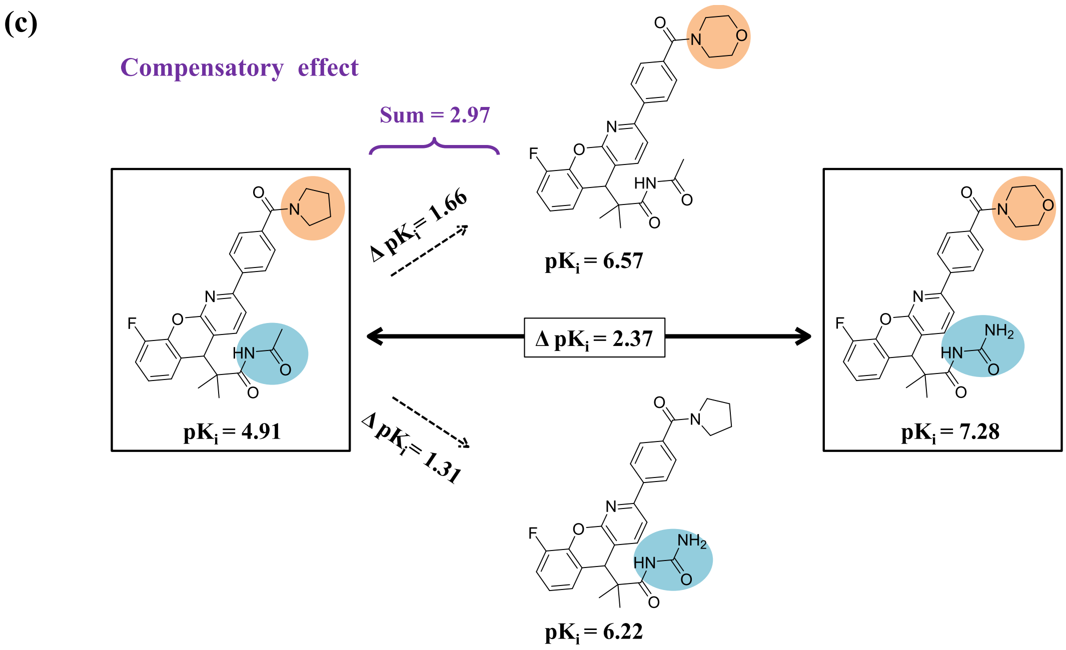


Fig. 6. (*continued*)

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bmc.2019.06.045.

References

1. Maggiora GM. *J Chem Inf Model.* 2006;46 1535 1535.
2. Stumpfe D, Bajorath J. *J Med Chem.* 2012;55:2932–2942.
3. Stumpfe D, Hu Y, Dimova D, Bajorath J. *J Med Chem.* 2014;57:18–28.
4. Bajorath J. *Future Med Chem.* 2014;6:1545–1549.
5. Kenny PW, Sadowski J. *Chemoinformatics in Drug Discovery.* Weinheim, Germany: Wiley-VCH; 2004:271–285.
6. Hussain J, Rea C. *J Chem Inf Model.* 2010;50:339–348.
7. Hu X, Hu Y, Vogt M, Stumpfe D, Bajorath J. *J Chem Inf Model.* 2012;52:1138–1145.
8. Lewell XQ, Judd DB, Watson SP, Hann MM. *J Chem Inf Comput Sci.* 1998;38:511–522.
9. de la Vega de León A, Bajorath J. *Med Chem Comm.* 2014;5:64–67.
10. Stumpfe D, Bajorath J. *Future Med Chem.* 2015;7:1565–1579.
11. Hu H, Stumpfe D, Bajorath J. *ACS Omega.* 2018;3:7736–7744.
12. Hu H, Stumpfe D, Bajorath J. *Future Med Chem.* 2019;11:379–394.
13. Stumpfe D, Dimova D, Bajorath J. *J Med Chem.* 2016;59:7667–7676.
14. Naveja JJ, Vogt M, Stumpfe D, Medina-Franco JL, Bajorath J. *ACS Omega.* 2019;4:1027–1032.
15. Gaulton A, Bellis LJ, Bento AP, et al. *Nucleic Acids Res.* 2012;40:D1100–D1107.
16. OEChem TK. *Version 1.7.7.* Santa Fe, NM, USA: OpenEye Scientific Software, Inc.; 2012.

Summary

In this analysis, we systematically explored ACs from ASs using high-confidence activity data from the ChEMBL database. Activity class-dependent potency difference criteria were calculated and applied to define ACs. The results indicated that more than 74% of ACs were formed by single-site analog pairs. ACs differing maximally five substitution sites were also observed, of which dual-site ACs dominated the distribution (91%). Since dual-site ACs carry two different R-groups at distinct substitution sites, we systematically identified single-site analogs that contained one of these two R-group replacements. The introduction of single-site analogs provided an opportunity to rationalize the formation of dual-site ACs. For only 297 dual-site ACs, single-site analogs for both substitution sites were found, hence characterizing the complete ACs. Potency comparisons between AC compounds and corresponding single-site analogs revealed that 141 ACs were redundant and could be represented by single-site ACs. For the remaining 156 confirmed dual-site ACs, three different potency effects for substituent combinations, i.e., synergistic, additive, and compensatory, were detected. Since potency effects in confirmed dual-site ACs were dominated by synergistic and compensatory effects, practical guidelines for compound optimization could be formulated.

Chapter 10

Conclusion

The “similarity-property principle" states that structurally similar compounds tend to display similar properties, e.g., biological activity. This principle depends heavily on how compound similarity is assessed. Given the rapidly increasing amount of compounds and activity data deposited in various databases, efficient computational methods for assessing compound similarity and performing large-scale SAR analysis are highly desirable. However, the presence of unexpected potency changes between structurally similar compounds forming ACs, indicates SAR discontinuity and often disappoints QSAR modeling. This does not imply, however, that AC formation should be negatively viewed. By contrast, if appropriately interpreted and rationalized, ACs are highly informative for SAR analysis. In this thesis, different approaches for addressing compound similarity are applied for AC analysis and associated informative AC data structures are introduced. Furthermore, alternative potency difference criteria for AC formation were investigated in detail.

Network representations have proven to be an indispensable tool to globally analyze and visualize AC information. However, increasing size and complexity of networks limit immediate SAR information assessment; thus, a new methodology aiming to simplify networks was derived based on a dual-round fragmentation scheme (*chapter 2*). The reduced network complemented the original AC network and simplified the analysis of cliff-associated SAR information. Given the popularity of the MMP concept, the MMP-cliff data structure was further extended through the inclusion of structural isomers, yielding isomer/MMP-cliffs (*chapter 3*). The introduction of isomer/MMP-cliffs indicated that the strategy of "walking" substituents at different core positions could be instructive for compound design. Then, the medicinal chemistry concept of privileged substructures (PSs) was explored using high-confidence activity data and related to AC analysis (*chapter 4*). Widespread distributions of PSs across different target families were observed indicating the promiscuous behavior of these PSs. PS-based AC network analysis implied that the structural context of PS embedding was critical for eliciting distinct biological activities. In order to increase the utilization of ACs in practice, a unified strategy for extracting dual-site ACs, isomer cliffs and PS-containing ACs was reported in *chapter 5*. Furthermore, for dual-site ACs, a four-compound data structure including two structural isomers was introduced to explore the influence of positional alternations on potency in detail. Beyond positional alterations of R-groups, single-atom modifications including atom replacements or atom walks represented minimal structural changes between compounds. The utilization of single-atom modifications in AC assessment resulted in the definition of atom-walk or atom-replacement ACs (*chapter 6*). The results indicated that around 3% of analog pairs with single-atom modifications formed ACs, the majority of which were atom-replacement ACs. Moreover, 59 ACs with at least one X-ray complex were available, which made the rationalization of AC formation at the structure-based level possible.

Large-scale AC analysis has been performed using curated data sets. Typically, only about 5% of analog pairs meet AC criteria. Some targets have a high propensity to form ACs while for others only limited numbers of ACs are available. In *chapter 7*, the frequency of occurrence of AC formation was rationalized by relating structural similarity relationships and potency distributions to each other. AC formation was found to be largely depended on potency fluctuations within RMMP-based clusters. To account for diverse potency distributions across different activity classes, activity class-depended potency difference criteria were introduced in *chapter 8*. In this study, statistically determined potency difference criteria were derived by considering potency value distributions of individual activity classes. Of note, activity class-dependent potency difference criteria were found to identify ACs that were more diversely distributed, emphasizing the presence of diverse structural contexts for AC analysis. In the final study (*chapter 9*), activity class-dependent AC criteria were applied to systematically analyze AC characteristics in analog series (ASs). Using AS membership as a criterion for structural similarity assessment, the overwhelming majority of ACs were found to be single-site ACs. Among ACs with multiple substitution sites, dual-site ACs (90.5%) dominated the distribution. Individual substitutions in dual-site ACs might lead to different potency effects (synergistic, compensatory, or additive). This finding implied that the systematic study of these effects, for instance, with the help of SAR matrices, can support practical applications in compound design.

In conclusion, this thesis introduced different approaches to assess molecular similarity and derive activity class-dependent potency difference criteria for AC assessment. The combinations of alternative similarity and potency difference criteria represent further evolution and refinement of the AC concept: from single- to multi-site ACs, from general to activity class-dependent ACs. Moreover, the incorporation of PSs, structural isomers, and single-atom modifications in AC analysis highlights practical relevance of AC analysis for medicinal chemistry.

Bibliography

[1] McKinney, J. D. The Practice of Structure Activity Relationships (SAR) in Toxicology. *Toxicological Sciences* **2000**, *56*, 8–17.

[2] Barbosa, F.; Horvath, D. Molecular Similarity and Property Similarity. *Current Topics in Medicinal Chemistry* **2004**, *4*, 589–600.

[3] Brown, A. C.; Fraser, T. R. On the Connection between Chemical Constitution and Physiological Action. Part I. On the Physiological Action of the Salts of the Ammonium Bases, Derived from Strychnia, Brucia,Thebaia, Codeia, Morphia, and Nicotia. *Philosophical Transactions of the Royal Society of London* **1868**, *25*, 151–203.

[4] 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 Research* **2012**, *40*, D1100–D1107.

[5] Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The Scaffold Tree - Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *Journal of Chemical Information and Modeling* **2007**, *47*, 47–58.

[6] Wassermann, A. M.; Haebel, P.; Weskamp, N.; Bajorath, J. SAR Matrices: Automated Extraction of Information-Rich SAR Tables from Large Compound Data Sets. *Journal of Chemical Information and Modeling* **2012**, *52*, 1769–1776.

[7] Peltason, L.; Bajorath, J. SAR Index: Quantifying the Nature of Structure-Activity Relationships. *Journal of Medicinal Chemistry* **2007**, *50*, 5571–5578.

[8] Méndez-Lucio, O.; Pérez-Villanueva, J.; Castillo, R.; Medina-Franco, J. L. Identifying Activity Cliff Generators of PPAR Ligands Using SAS Maps. *Molecular Informatics* **2012**, *31*, 837–846.

[9] 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. *Journal of Chemical Information and Modeling* **2010**, *50*, 1021–1033.

[10] Wawer, M.; Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Structure-Activity Relationship Anatomy by Network-like Similarity Graphs and Local Structure-Activity Relationship Indices. *Journal of Medicinal Chemistry* **2008**, *51*, 6075–6084.

[11] Gupta-Ostermann, D.; Hu, Y.; Bajorath, J. Introducing the LASSO Graph for Compound Data Set Representation and Structure–Activity Relationship Analysis. *Journal of Medicinal Chemistry* **2012**, *55*, 5546–5553.

[12] Zhang, B.; Hu, Y.; Bajorath, J. AnalogExplorer: a New Method for Graphical Analysis of Analog Series and Associated Structure–Activity Relationship Information. *Journal of Medicinal Chemistry* **2014**, *57*, 9184–9194.

[13] Isarankura-Na-Ayudhya, C.; Naenna, T.; Nantasenamat, C.; Prachayasittikul, V. A Practical Overview of Quantitative Structure-Activity Relationship. *EXCLI Journal* **2009**, *8*, 74–88.

[14] Yousefinejad, S.; Hemmateenejad, B. Chemometrics Tools in QSAR/QSPR Studies: a historical perspective. *Chemometrics and Intelligent Laboratory Systems* **2015**, *149*, 177–204.

[15] Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity Prediction using Deep Learning. *Frontiers in Environmental Science* **2016**, *3*, 80.

[16] Yang, Q.; Li, Y.; Yang, J.; Liu, Y.; Zhang, L.; Luo, S.; Cheng, J. Holistic Prediction of the pK_a in Diverse Solvents Based on a Machine-Learning Approach. *Angewandte Chemie International Edition* **2020**, *59*, 19282–19291.

[17] Plante, J.; Werner, S. JPlogP: An Improved logP Predictor Trained Using Predicted Data. *Journal of Cheminformatics* **2018**, *10*, e61.

[18] Miljković, F.; Rodríguez-Pérez, R.; Bajorath, J. Machine Learning Models for Accurate Prediction of Kinase Inhibitors with Different Binding Modes. *Journal of Medicinal Chemistry* **2020**, *63*, 8738–8748.

[19] Johnson, M.; Maggiora, G.; Meeting, A. C. S. *Concepts and Applications of Molecular Similarity*; A Wiley-Interscience publication; Wiley, 1990.

[20] Grover, M.; Singh, B.; Bakshi, M.; Singh, S. Quantitative Structure–Property Relationships in Pharmaceutical Research – Part 1. *Pharmaceutical Science & Technology Today* **2000**, *3*, 28–35.

[21] Grover, M.; Singh, B.; Bakshi, M.; Singh, S. Quantitative Structure–Property Relationships in Pharmaceutical Research – Part 2. *Pharmaceutical Science & Technology Today* **2000**, *3*, 50–57.

[22] Muratov, E. N. et al. QSAR without Borders. *Chemical Society Reviews* **2020**, *49*, 3525–3564.

[23] Maggiora, G. M. On Outliers and Activity Cliffs-Why QSAR Often Disappoints. *Journal of Chemical Information and Modeling* **2006**, *46*, 1535–1535.

[24] Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. *Journal of Medicinal Chemistry* **2013**, *57*, 18–28.

[25] Bajorath, J. Duality of Activity Cliffs in Drug Discovery. *Expert Opinion on Drug Discovery* **2019**, *14*, 517–520.

[26] Cruz-Monteagudo, M.; Medina-Franco, J. L.; Pérez-Castillo, Y.; Nicolotti, O.; Cordeiro, M. N. D.; Borges, F. Activity Cliffs in Drug Discovery: Dr Jekyll or Mr Hyde? *Drug Discovery Today* **2014**, *19*, 1069–1080.

[27] de la Vega de León, A.; Bajorath, J. Formation of Activity Cliffs Is Accompanied by Systematic Increases in Ligand Efficiency from Lowly to Highly Potent Compounds. *The AAPS Journal* **2014**, *16*, 335–341.

[28] Hopkins, A. L.; Keserü, G. M.; Leeson, P. D.; Rees, D. C.; Reynolds, C. H. The Role of Ligand Efficiency Metrics in Drug Discovery. *Nature Reviews Drug Discovery* **2014**, *13*, 105–121.

[29] David, L.; Thakkar, A.; Mercado, R.; Engkvist, O. Molecular Representations in AI-Driven Drug Discovery: a Review and Practical Guide. *Journal of Cheminformatics* **2020**, *12*, e56.

[30] Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Modeling* **1988**, *28*, 31–36.

[31] Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry. *Journal of Medicinal Chemistry* **2013**, *57*, 3186–3204.

[32] Chuang, K. V.; Gunsalus, L. M.; Keiser, M. J. Learning Molecular Representations for Medicinal Chemistry. *Journal of Medicinal Chemistry* **2020**, *63*, 8705–8722.

[33] *MACCS Structural Keys*; Accelrys: San Diego, CA, 2011.

[34] Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50*, 742–754.

[35] Schwaha, R. The Similarity Principle – New Trends and Applications in Ligand-Based Drug Discovery and ADMET Profiling. *Scientia Pharmaceutica* **2008**, *76*, 5–18.

[36] Willett, P. *Methods in Molecular Biology*; Humana Press, 2010; pp 133–158.

[37] Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *Journal of Medicinal Chemistry* **2012**, *55*, 2932–2942.

[38] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research* **2000**, *28*, 235–242.

[39] Peltason, L.; Bajorath, J. Molecular Similarity Analysis Uncovers Heterogeneous Structure-Activity Relationships and Variable Activity Landscapes. *Chemistry & Biology* **2007**, *14*, 489–497.

[40] Hu, Y.; Furtmann, N.; Gütschow, M.; Bajorath, J. Systematic Identification and Classification of Three-Dimensional Activity Cliffs. *Journal of Chemical Information and Modeling* **2012**, *52*, 1490–1498.

[41] Chen, Y.; Mathai, N.; Kirchmair, J. Scope of 3D Shape-Based Approaches in Predicting the Macromolecular Targets of Structurally Complex Small Molecules Including Natural Products and Macrocyclic Ligands. *Journal of Chemical Information and Modeling* **2020**, *60*, 2858–2875.

[42] Koes, D. R.; Camacho, C. J. Shape-Based Virtual Screening with Volumetric Aligned Molecular Shapes. *Journal of Computational Chemistry* **2014**, *35*, 1824–1834.

[43] Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. Matched Molecular Pairs as a Medicinal Chemistry Tool. *Journal of Medicinal Chemistry* **2011**, *54*, 7739–7750.

[44] Stumpfe, D.; Dimova, D.; Bajorath, J. Computational Method for the Systematic Identification of Analog Series and Key Compounds Representing Series and Their Biological Activity Profiles. *Journal of Medicinal Chemistry* **2016**, *59*, 7667–7676.

[45] Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J. MMP-Cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs. *Journal of Chemical Information and Modeling* **2012**, *52*, 1138–1145.

[46] Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *Journal of Chemical Information and Modeling* **2010**, *50*, 339–348.

[47] Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP-Retrosynthetic Combinatorial Analysis Procedure: a Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *Journal of Chemical Information and Computer Sciences* **1998**, *38*, 511–522.

[48] de la Vega de León, A.; Bajorath, J. Matched Molecular Pairs Derived by Retrosynthetic Fragmentation. *MedChemCommun* **2014**, *5*, 64–67.

[49] O'Boyle, N. M.; Boström, J.; Sayle, R. A.; Gill, A. Using Matched Molecular Series as a Predictive Tool To Optimize Biological Activity. *Journal of Medicinal Chemistry* **2014**, *57*, 2704–2713.

[50] Dimova, D.; Bajorath, J. Extraction of SAR Information From Activity Cliff Clusters via Matching Molecular Series. *European Journal of Medicinal Chemistry* **2014**, *87*, 454–460.

[51] Schuetz, D. A.; Richter, L.; Martini, R.; Ecker, G. F. A Structure–Kinetic Relationship Study Using Matched Molecular Pair Analysis. *RSC Medicinal Chemistry* **2020**, *11*, 1285–1294.

[52] Naveja, J. J.; Vogt, M.; Stumpfe, D.; Medina-Franco, J. L.; Bajorath, J. Systematic Extraction of Analogue Series from Large Compound Collections Using a New Computational Compound–Core Relationship Method. *ACS Omega* **2019**, *4*, 1027–1032.

[53] Schuffenhauer, A.; Varin, T. Rule-Based Classification of Chemical Structures by Scaffold. *Molecular Informatics* **2011**, *30*, 646–664.

[54] Hu, Y.; Stumpfe, D.; Bajorath, J. Lessons Learned from Molecular Scaffold Analysis. *Journal of Chemical Information and Modeling* **2011**, *51*, 1742–1753.

[55] Hu, Y.; Stumpfe, D.; Bajorath, J. Computational Exploration of Molecular Scaffolds in Medicinal Chemistry. *Journal of Medicinal Chemistry* **2016**, *59*, 4062–4076.

[56] Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of Medicinal Chemistry* **1996**, *39*, 2887–2893.

[57] Xu, Y.; Johnson, M. Algorithm for Naming Molecular Equivalence Classes Represented by Labeled Pseudographs. *Journal of Chemical Information and Computer Sciences* **2001**, *41*, 181–185.

[58] Dimova, D.; Bajorath, J. Is Scaffold Hopping a Reliable Indicator for the Ability of Computational Methods to Identify Structurally Diverse Active Compounds? *Journal of Computer-Aided Molecular Design* **2017**, *31*, 603–608.

[59] Dimova, D.; Stumpfe, D.; Hu, Y.; Bajorath, J. Analog Series-Based Scaffolds: Computational Design and Exploration of a New Type of Molecular Scaffolds for Medicinal Chemistry. *Future Science OA* **2016**, *2*, FSO149.

[60] Dimova, D.; Stumpfe, D.; Bajorath, J. Computational Design of New Molecular Scaffolds for Medicinal Chemistry, Part II: Generalization of Analog Series-Based Scaffolds. *Future Science OA* **2018**, *4*, FSO267.

[61] Dimova, D.; Bajorath, J. Collection of Analog Series-Based Scaffolds from Public Compound Sources. *Future Science OA* **2018**, *4*, FSO287.

[62] Hu, Y.; Stumpfe, D.; Bajorath, J. Computational Exploration of Molecular Scaffolds in Medicinal Chemistry. *Journal of Medicinal Chemistry* **2016**, *59*, 4062–4076.

[63] Evans, B. E. et al. Methods for Drug Discovery: Development of Potent, Selective, Orally Effective Cholecystokinin Antagonists. *Journal of Medicinal Chemistry* **1988**, *31*, 2235–2246.

[64] Welsch, M. E.; Snyder, S. A.; Stockwell, B. R. Privileged Scaffolds for Library Design and Drug Discovery. *Current Opinion in Chemical Biology* **2010**, *14*, 347–361.

[65] Schnur, D. M.; Hermsmeier, M. A.; Tebben, A. J. Are Target-Family-Privileged Substructures Truly Privileged? *Journal of Medicinal Chemistry* **2006**, *49*, 2000–2009.

[66] Severinsen, R.; Bourne, G. T.; Tran, T. T.; Ankersen, M.; Begtrup, M.; Smythe, M. L. Library of Biphenyl Privileged Substructures Using a Safety-Catch Linker Approach. *Journal of Combinatorial Chemistry* **2008**, *10*, 557–566.

[67] Wan, Y.; Li, Y.; Yan, C.; Yan, M.; Tang, Z. Indole: a Privileged Scaffold for the Design of Anti-Cancer Agents. *European Journal of Medicinal Chemistry* **2019**, *183*, 111691.

[68] Thanikachalam, P. V.; Maurya, R. K.; Garg, V.; Monga, V. An Insight into the Medicinal Perspective of Synthetic Analogs of Indole: a Review. *European Journal of Medicinal Chemistry* **2019**, *180*, 562–612.

[69] Stumpfe, D.; Bajorath, J. Assessing the Confidence Level of Public Domain Compound Activity Data and the Impact of Alternative Potency Measurements on SAR Analysis. *Journal of Chemical Information and Modeling* **2011**, *51*, 3131–3137.

[70] Dimova, D.; Bajorath, J. Advances in Activity Cliff Research. *Molecular Informatics* **2016**, *35*, 181–191.

[71] Hu, Y.; Stumpfe, D.; Bajorath, J. Advancing the Activity Cliff Concept. *F1000Research* **2013**, *2*, 199.

[72] Hu, Y.; Bajorath, J. Extending the Activity Cliff Concept: Structural Categorization of Activity Cliffs and Systematic Identification of Different Types of Cliffs in the ChEMBL Database. *Journal of Chemical Information and Modeling* **2012**, *52*, 1806–1811.

[73] Schneider, N.; Lewis, R. A.; Fechner, N.; Ertl, P. Chiral Cliffs: Investigating the Influence of Chirality on Binding Affinity. *ChemMedChem* **2018**, *13*, 1315–1324.

[74] Pérez-Benito, L.; Casajuana-Martin, N.; Jiménez-Rosés, M.; van Vlijmen, H.; Tresadern, G. Predicting Activity Cliffs with Free-Energy Perturbation. *Journal of Chemical Theory and Computation* **2019**, *15*, 1884–1895.

[75] Horvath, D.; Marcou, G.; Varnek, A.; Kayastha, S.; de la Vega de León, A.; Bajorath, J. Prediction of Activity Cliffs Using Condensed Graphs of Reaction Representations, Descriptor Recombination, Support Vector Machine Classification, and Support Vector Regression. *Journal of Chemical Information and Modeling* **2016**, *56*, 1631–1640.

[76] Husby, J.; Bottegoni, G.; Kufareva, I.; Abagyan, R.; Cavalli, A. Structure-Based Predictions of Activity Cliffs. *Journal of Chemical Information and Modeling* **2015**, *55*, 1062–1076.

[77] Walters, W. P.; Namchuk, M. Designing Screens: How to Make Your Hits a Hit. *Nature Reviews Drug Discovery* **2003**, *2*, 259–266.

[78] Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V. S.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U.; Sittampalam, G. S. Impact of High-Throughput Screening in Biomedical Research. *Nature Reviews Drug Discovery* **2011**, *10*, 188–195.

[79] Jasial, S.; Hu, Y.; Bajorath, J. Determining the Degree of Promiscuity of Extensively Assayed Compounds. *Plos One* **2016**, *11*, e0153873.

[80] Wassermann, A. M. et al. Dark Chemical Matter as a Promising Starting Point for Drug Lead Discovery. *Nature Chemical Biology* **2015**, *11*, 958–966.

[81] Jasial, S.; Bajorath, J. Dark Chemical Matter in Public Screening Assays and Derivation of Target Hypotheses. *MedChemComm* **2017**, *8*, 2100–2104.

[82] Ballante, F.; Rudling, A.; Zeifman, A.; Luttens, A.; Vo, D. D.; Irwin, J. J.; Kihlberg, J.; Brea, J.; Loza, M. I.; Carlsson, J. Docking Finds GPCR Ligands in Dark Chemical Matter. *Journal of Medicinal Chemistry* **2019**, *63*, 613–620.

[83] Hu, Y.; Maggiora, G. M.; Bajorath, J. Activity Cliffs in PubChem Confirmatory Bioassays Taking Inactive Compounds into Account. *Journal of Computer-Aided Molecular Design* **2013**, *27*, 115–124.

[84] Stumpfe, D.; Bajorath, J. Frequency of Occurrence and Potency Range Distribution of Activity Cliffs in Bioactive Compounds. *Journal of Chemical Information and Modeling* **2012**, *52*, 2348–2353.

[85] Stumpfe, D.; Bajorath, J. Monitoring Global Growth of Activity Cliff Information over Time and Assessing Activity Cliff Frequencies and Distributions. *Future Medicinal Chemistry* **2015**, *7*, 1565–1579.

[86] Stumpfe, D.; Dimova, D.; Bajorath, J. Composition and Topology of Activity Cliff Clusters Formed by Bioactive Compounds. *Journal of Chemical Information and Modeling* **2014**, *54*, 451–461.

[87] Vogt, M.; Huang, Y.; Bajorath, J. From Activity Cliffs to Activity Ridges: Informative Data Structures for SAR Analysis. *Journal of Chemical Information and Modeling* **2011**, *51*, 1848–1856.

[88] Dimova, D.; Heikamp, K.; Stumpfe, D.; Bajorath, J. Do Medicinal Chemists Learn from Activity Cliffs? A Systematic Evaluation of Cliff Progression in Evolving Compound Data Sets. *Journal of Medicinal Chemistry* **2013**, *56*, 3339–3345.

[89] Dimova, D.; Stumpfe, D.; Bajorath, J. Systematic Assessment of Coordinated Activity Cliffs Formed by Kinase Inhibitors and Detailed Characterization of Activity Cliff Clusters and Associated SAR Information. *European Journal of Medicinal Chemistry* **2015**, *90*, 414–427.

[90] Dimova, D.; Stumpfe, D.; Bajorath, J. Method for the Evaluation of Structure–Activity Relationship Information Associated with Coordinated Activity Cliffs. *Journal of Medicinal Chemistry* **2014**, *57*, 6553–6563.

[91] Dimova, D.; Stumpfe, D.; Hu, Y.; Bajorath, J. Activity Cliff Clusters as a Source of Structure–Activity Relationship Information. *Expert Opinion on Drug Discovery* **2015**, *10*, 441–447.

[92] Wassermann, A. M.; Bajorath, J. A Data Mining Method to Facilitate SAR Transfer. *Journal of Chemical Information and Modeling* **2011**, *51*, 1857–1866.

[93] Furtmann, N.; Hu, Y.; Gütschow, M.; Bajorath, J. Identification of Interaction Hot Spots in Structures of Drug Targets on the Basis of Three-Dimensional Activity Cliff Information. *Chemical Biology & Drug Design* **2015**, *86*, 1458–1465.

[94] Furtmann, N.; Hu, Y.; Gütschow, M.; Bajorath, J. Identification and Analysis of the Currently Available High-Confidence Three-Dimensional Activity Cliffs. *RSC Advances* **2015**, *5*, 43660–43668.

[95] Bissantz, C.; Kuhn, B.; Stahl, M. A Medicinal Chemist's Guide to Molecular Interactions. *Journal of Medicinal Chemistry* **2010**, *53*, 5061–5084.

[96] Deng, Z.; Chuaqui, C.; Singh, J. Structural Interaction Fingerprint (SIFt): a Novel Method for Analyzing Three-Dimensional Protein-Ligand Binding Interactions. *Journal of Medicinal Chemistry* **2004**, *47*, 337–344.

[97] Vass, M.; Kooistra, A. J.; Ritschel, T.; Leurs, R.; de Esch, I. J.; de Graaf, C. Molecular Interaction Fingerprint Approaches for GPCR Drug Discovery. *Current Opinion in Pharmacology* **2016**, *30*, 59–68.

[98] Singh, J.; Deng, Z.; Narale, G.; Chuaqui, C. Structural Interaction Fingerprints: a New Approach to Organizing, Mining, Analyzing, and Designing Protein-Small Molecule Complexes. *Chemical Biology & Drug Design* **2006**, *67*, 5–12.

[99] Da, C.; Stashko, M.; Jayakody, C.; Wang, X.; Janzen, W.; Frye, S.; Kireev, D. Discovery of MER Kinase Inhibitors by Virtual Screening Using Structural Protein–Ligand Interaction Fingerprints. *Bioorganic & Medicinal Chemistry* **2015**, *23*, 1096–1101.

[100] Méndez-Lucio, O.; Kooistra, A. J.; de Graaf, C.; Bender, A.; Medina-Franco, J. L. Analyzing Multitarget Activity Landscapes Using Protein–Ligand Interaction Fingerprints: Interaction Cliffs. *Journal of Chemical Information and Modeling* **2015**, *55*, 251–262.

[101] Rodríguez-Pérez, R.; Miljković, F.; Bajorath, J. Assessing the Information Content of Structural and Protein–Ligand Interaction Representations for the Classification of Kinase Inhibitor Binding Modes via Machine Learning and Active Learning. *Journal of Cheminformatics* **2020**, *12*, e36.

[102] Abramyan, T. M.; An, Y.; Kireev, D. Off-Pocket Activity Cliffs: a Puzzling Facet of Molecular Recognition. *Journal of Chemical Information and Modeling* **2020**, *60*, 152–161.

[103] Schönherr, H.; Cernak, T. Profound Methyl Effects in Drug Discovery and a Call for New C-H Methylation Reactions. *Angewandte Chemie International Edition* **2013**, *52*, 12256–12267.

[104] Babine, R. E.; Bender, S. L. Molecular Recognition of Protein-Ligand Complexes: Applications to Drug Design. *Chemical Reviews* **1997**, *97*, 1359–1472.

[105] Caro, J. A.; Harpole, K. W.; Kasinath, V.; Lim, J.; Granja, J.; Valentine, K. G.; Sharp, K. A.; Wand, A. J. Entropy in Molecular Recognition by Proteins. *Proceedings of the National Academy of Sciences* **2017**, *114*, 6563–6568.

[106] de Freitas, R. F.; Schapira, M. A Systematic Analysis of Atomic Protein–Ligand Interactions in the PDB. *MedChemComm* **2017**, *8*, 1970–1981.

[107] Alfonso-Prieto, M.; Navarini, L.; Carloni, P. Understanding Ligand Binding to G-Protein Coupled Receptors Using Multiscale Simulations. *Frontiers in Molecular Biosciences* **2019**, *6*, 29.

[108] Hollingsworth, S. A.; Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **2018**, *99*, 1129–1143.

Additional Publications

Hu, H.; Laufkötter, O.; Miljković, F.; Bajorath, J. Systematic Comparison of Competitive and Allosteric Kinase Inhibitors Reveals Common Structural Characteristics. *European Journal of Medicinal Chemistry* **2021**, *214*, 113206.

Laufkötter, O.; Hu, H.; Miljković, F.; Bajorath, J. Structure- and Similarity-Based Survey of Allosteric Kinase Inhibitors and Activators and Closely Related Compounds. *Journal of Medicinal Chemistry* **2021**, https://doi.org/10.1021/acs.jmedchem.0c02076.

Stumpfe, D.; Hu, H.; Bajorath, J. Computational Method for the Identification of Third Generation Activity Cliffs. *MethodsX* **2020**, *7*, e100793.

Hu, H.; Bajorath, J. Exploring Structure-Promiscuity Relationships Using Dual-Site Promiscuity Cliffs and Corresponding Single-Site Analogs. *Bioorganic & Medicinal Chemistry*, **2020**, *28*, 115238.

Stumpfe, D.; Hu, H.; Bajorath, J. Advances in Exploring Activity Cliffs. *Journal of Computer-Aided Molecular Design* **2020**, *34*, 929-942.

Hu, H.; Bajorath, J. Evidence for the Presence of Core Structure-Dependent Activity Cliffs. *Future Medicinal Chemistry* **2020**, *12*, 1451-1455.

Hu, H.; Bajorath, J. Data Set of Activity Cliffs with Single-Atom Modification and Associated X-Ray Structure Information for Medicinal and Computational Chemistry Applications. *Data in Brief* **2020**, *33*, 106364.

Hu, H.; Stumpfe, D.; Bajorath, J. Systematic Identification of Target Set-Dependent Activity Cliffs. *Future Science OA* **2019**, *5*, FSO363.

Stumpfe, D.; Hu, H.; Bajorath, J. The Evolving Concept of Activity Cliffs. *ACS Omega* **2019**, *4*, 14360-14368.