

Computational Methods Generating High-Resolution Views of Complex Structure-Activity Relationships

Kumulative Dissertation

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

der Mathematisch-Naturwissenschaftlichen Fakultät

der Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

DILYANA KALINOVA DIMOVA

aus Burgas, Bulgarien

Bonn

February, 2014

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: 23.04.2014
Erscheinungsjahr: 2014

Abstract

The analysis of structure-activity relationships (SARs) of small bioactive compounds is a central task in medicinal chemistry and pharmaceutical research. The study of SARs is in principle not limited to computational methods, however, as data sets rapidly grow in size, advanced computational approaches become indispensable for SAR analysis. Activity landscapes are one of the preferred and widely used computational models to study large-scale SARs. Activity cliffs are cardinal features of activity landscape representations and are thought to contain high SAR information content.

This work addresses major challenges in systematic SAR exploration and specifically focuses on the design of novel activity landscape models and comprehensive activity cliff analysis. In the first part of the thesis, two conceptually different activity landscape representations are introduced for compounds active against multiple targets. These models are designed to provide an intuitive graphical access to compounds forming single and multi-target activity cliffs and displaying multi-target SAR characteristics. Further, a systematic analysis of the frequency and distribution of activity cliffs is carried out. In addition, a large-scale data mining effort is designed to quantify and analyze fingerprint-dependent changes in SAR information. The second part of this work is dedicated to the concept of activity cliffs and their utility in the practice of medicinal chemistry. Therefore, a computational approach is introduced to search for detectable SAR advantages associated with activity cliffs. In addition, the question is investigated to what extent activity cliffs might be utilized as starting points in practical compound optimization efforts. Finally, all activity cliff configurations formed by currently available bioactive compounds are thoroughly examined. These configurations are further classified and their frequency of occurrence and target distribution are determined. Furthermore, the activity cliff concept is extended to explore the relation between chemical structures and compound promiscuity. The notion of promiscuity cliffs is introduced to deduce structural modifications that might induce large-magnitude promiscuity effects.

To my beloved family.

Acknowledgments

I would like to first express my gratitude to my supervisor Prof. Dr. Jürgen Bajorath for introducing me to the exciting subject of Chemoinformatics, for his continuous scientific inspiration and personal support, and devoted guidance during my doctoral studies. I also thank Prof. Dr. Michael Gütschow for reviewing my thesis as a co-referent.

I further thank to all my colleagues of the LSI group for the fruitful scientific discussions and interactive working atmosphere. I would particularly like to thank my dear friends Kathrin Heikamp, Ye Hu, Dagmar Stumpfe, and Anne Wassermann for the good times we shared, for the productive late-night brainstorming sessions and colorful whiteboard presentations. Special thanks to all regulars' table members for the entertaining time spent together.

Finally, I would like to express my love and gratitude to my family for being the greatest inspiration, for their persistent moral support and for everything they have ever done for me during all the years in Germany.

Contents

1 Introduction	1
Molecular Similarity	2
SAR Exploration	9
Activity Landscapes	10
Activity Cliffs	20
Thesis Outline	23
References	27
2 Design of Multi-Target Activity Landscapes That Capture Hierarchical Activity Cliff Distributions	29
Introduction	29
Publication	31
Summary	41
3 Navigating High-Dimensional Activity Landscapes: Design and Application of the Ligand-Target Differentiation Map	43
Introduction	43
Publication	45
Summary	53
4 Comprehensive Analysis of Single- and Multi-Target Activity Cliffs Formed by Currently Available Bioactive Compounds	55
Introduction	55
Publication	57

Summary	63
5 Quantifying the Fingerprint Descriptor Dependence of Structure-Activity Relationship Information on a Large Scale	65
Introduction	65
Publication	67
Summary	75
6 Compound Pathway Model To Capture SAR Progression: Comparison of Activity Cliff-Dependent and -Independent Pathways	77
Introduction	77
Publication	79
Summary	85
7 Do Medicinal Chemists Learn from Activity Cliffs? A Systematic Evaluation of Cliff Progression in Evolving Compound Data Sets	87
Introduction	87
Publication	89
Summary	97
8 Composition and Topology of Activity Cliff Clusters Formed by Bioactive Compounds	99
Introduction	99
Publication	101
Summary	111
9 Matched Molecular Pair Analysis of Small Molecule Microarray Data Identifies Promiscuity Cliffs and Reveals Molecular Origins of Extreme Compound Promiscuity	113
Introduction	113
Publication	115
Summary	125

CONTENTS

10 Conclusions	127
Additional References	131
Additional Publications	133

Chapter 1

Introduction

Over the past decades, the study of small bioactive molecules and their interactions with biological targets has played, and continues to play, a central role in elucidating biological processes and understanding protein functions. To these ends, understanding the relation between the chemical structures and the biological activity of active compounds, commonly referred to as *structure-activity relationships* (SARs), is a primary task in medicinal chemistry and pharmaceutical research.¹ The fundamental goal of SAR analysis is to demonstrate how structural changes might affect the biological activity of compounds, and further identify structural modifications which translate into compound potency improvement. Supported by a wealth of observations, the SARs are often in accord with the similarity-property principle (SPP)² - a central paradigm in medicinal chemistry, stating that similar molecules should exhibit similar biological functions. However, it is also well-appreciated that exceptions do exist, and that structurally analogous compounds may display different SAR characteristics. For example, small structural modifications can dramatically change the biological activity, thereby significantly increasing or decreasing the compound activity.³ These considerations have demonstrated that SARs are multi-faceted in nature, an observation that still greatly challenges the SAR exploration and makes it a highly sophisticated task.

Molecular Similarity

To assess the relationship between structural modifications and biological activity, molecules must be represented in a consistent and well-defined manner. In addition, to compare changes in biological responses, potency annotations for the underlying compounds must be provided. Furthermore, mostly, but not exclusively, the SAR analysis of compound data sets frequently relies on pairwise structural comparisons of small molecules. To these ends, the application of similarity measures that quantify the degree of structural relatedness between compounds becomes indispensable for the SAR analysis.

The assessment of structural similarity of compounds can be regarded as a two-step procedure. First, a molecular representation is chosen that encodes relevant molecular and/or chemical features. A similarity metric, often termed a similarity coefficient, is then used to quantitatively evaluate the molecular similarity on the basis of the chosen molecular representation. Hence, the outcome of similarity evaluation might substantially be influenced by the chosen molecular representation and similarity metric.^{4,5}

Molecular Representations

A variety of molecular representations have been introduced thus far.^{6,7} In general, representations can be subdivided into three different categories: one-, two-, and three-dimensional (1D, 2D, and 3D, respectively). Examples for 1D-representations include simple chemical composition formula and more complex notations, such as the SMILES⁸ language that serves as a universal chemical nomenclature to represent chemical structure information.

One of the best known and most widely used representations of small molecules is the molecular graph (Figure 1.1). These 2D graphs can be considered as the “natural language of medicinal chemists” and serve as simplified and intuitive models of molecular structures. In these graphs, nodes denote atoms using atomic symbols and edges encode bonding information. Therefore, 2D molecular graphs represent the connectivity between atoms and the topology of the molecules.

To account for conformational information, 3D representations based on the spacial arrangements of the atoms have been introduced. Notable examples include molecular surfaces and pharmacophore models. A pharmacophore is defined as the 3D arrangement of atoms, groups, or functions that is essential for a molecule to specifically interact with a biological target.

On the basis of different molecular representations, mathematical models have been introduced to capture a variety of chemical properties and are commonly termed molecular descriptors. Examples of simple descriptors include molecular weight, number of heavy atoms and number of aromatic rings. Similar to the molecular representations, descriptors are classified as 1D, 2D, or 3D, depending on the dimensionality of the utilized representation.

Molecular fingerprints are a special kind of descriptors that are used to characterize chemical structure and/or molecular properties of a molecule. In chemoinformatics and pharmaceutical research, fingerprints are generally defined as bit string representations. Over the past years, a variety of fingerprints have been introduced that considerably differ in their design, composition, and complexity.⁹ Although fingerprints are string representations, and hence one-dimensional, they are typically classified as 2D (i.e., based on molecular graphs) and 3D (i.e., based on molecular conformations) fingerprints.

Usually, but not exclusively, fingerprints are in a binary format, i.e., each bit position accounts for the presence or absence of a given feature. If the feature is present, the corresponding bit is set to 1, otherwise it is set to 0. In addition, fingerprints are mostly of a fixed length.

Substructure fingerprints are one of the major prototypes of 2D fingerprints and can be considered as dictionaries of predefined substructures. Classical examples for substructure fingerprints (also termed keyed fingerprints) include the Molecular ACCess System (MACCS)¹⁰ structural keys consisting of 166 structural features each corresponding to a specific bit position. An example of a keyed fingerprint is shown in Figure 1.1a in which features present in the molecule are colored in gray. The corresponding bits are set to 1 (gray shades) in the representation. Non-binary versions of fingerprints, also termed count fingerprints, have also been developed. Here, each position numerically accounts for the frequency of occurrence of the underlying feature.^{11,12}

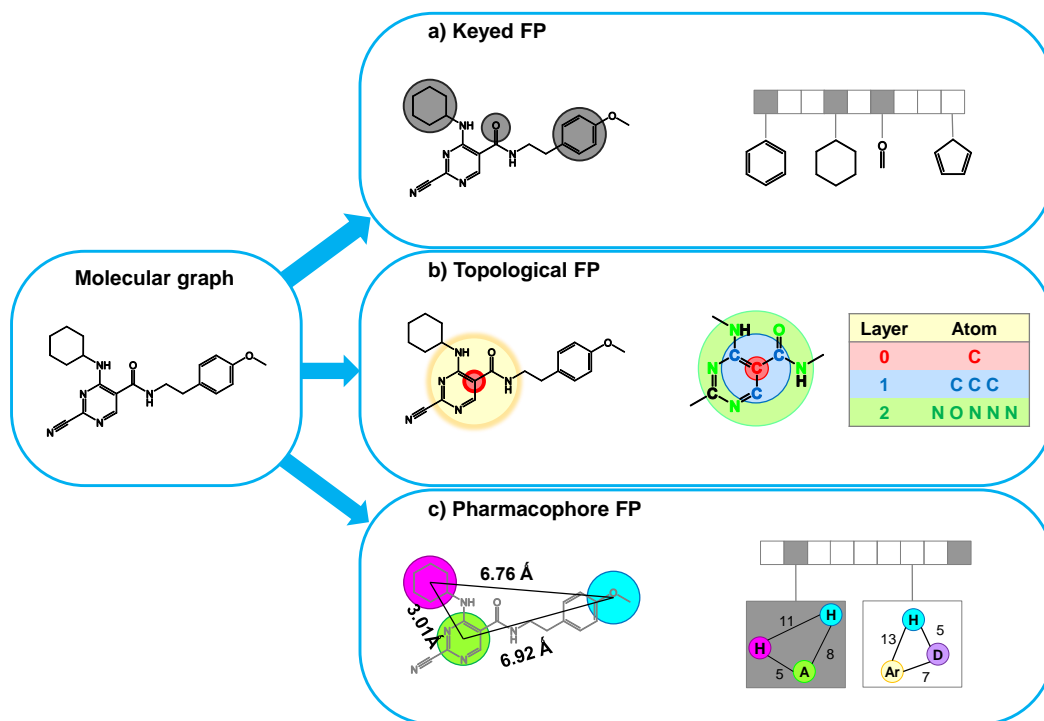


Figure 1.1: Molecular fingerprints. Three different types of molecular fingerprints are shown. Specific molecular structure information used to derive the corresponding fingerprint representation is highlighted. Adapted from [13].

On the basis of molecular topologies, topological fingerprints have been introduced and account for connectivity pathways between atoms in a molecule. Representative examples include the Daylight fingerprint often consisting of 2048 bits and the MOLPRINT 2D^{14,15} fingerprint that has a variable length. In Daylight fingerprint, paths through the molecule are calculated until a predefined length is reached (bond distance). By applying a hashing function, these paths are subsequently mapped onto a string of a fixed length. In contrast to keyed fingerprints, individual bit positions in hashed fingerprints do not correspond to individual structural features and hence, cannot be chemically interpreted. Different from the Daylight fingerprint, the ECFP4 fingerprint¹⁶ is designed to capture connectivity information in layered atom environments with a maximum diameter of four bonds around each atom. These calculations are molecule-specific resulting in a fingerprint of variable length. The atom

environment perception of a topological fingerprint of bond diameter four is schematically illustrated on Figure 1.1b.

Pharmacophore fingerprints capture pharmacophore patterns. Examples include the Typed Graph Triangle (TGT)¹⁷ and Typed Graph Distance (TGD)¹⁷ fingerprints consisting of 1704 and 420 bit positions, respectively, which are Molecular Operating Environment (MOE)¹⁷ internal developments. In the TGD fingerprint, shortest distances (in terms of number of bonds) in the molecular graph between two atoms (represented as seven pharmacophore features) are calculated and assigned to 15 distance ranges to monitor distances between feature pairs. In contrast, the TGT fingerprint is designed to capture three-point pharmacophore patterns in molecular graphs. Atoms are assigned to one of four different atom types (hydrogen-bond donor, hydrogen-bond acceptor, donor/acceptor, or hydrophobic). Applied are graph (bond) distances subdivided into six distance ranges. Exemplary 2D pharmacophore pattern information encoded in a pharmacophore-based fingerprint is highlighted in Figure 1.1c.

Similarity Coefficients

As stated above, similarity coefficients are applied to account for the degree of similarity between compounds. Although a wide-range of coefficients and distance functions have been introduced, the most widely used is the Jaccard or Tanimoto coefficient (Tc).^{6,18} For two fingerprints A and B , the Tanimoto coefficient calculates the ratio of the number of bits set on in both fingerprints over the number of bits set on in either fingerprint. Formally, the Tc is defined as follows:

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

where a and b denote the number of bits set on in fingerprint A and B , respectively, whereas c denotes the number of bits set on in both fingerprints. The Tc ranges between 0 and 1, with 0 corresponding to no fingerprint overlap and 1 to identical fingerprints. It should be noted that, identical fingerprints do not nec-

essarily correspond to identical molecules (as fingerprints are only abstractions of molecular structures). Furthermore, as defined by the above formula, the Tc only takes into account bits set to 1 (i.e., features present in the molecule). Hence, the magnitude of the Tc value will be greatly influenced by the bit density in the underlying fingerprint, which on the other hand, increases with molecular size and complexity.¹⁹

The calculation of Tc translates structural similarity into numerical values and can be interpreted as the “percentage of structural features shared between two compounds”, yet it is debatable which Tc value corresponds to “significant similarity”. There is no generally applicable Tc threshold for the indication of structural similarity, which is dependent on the molecular fingerprint applied.²⁰ However, for SAR applications, a threshold value of 55% and 85% are typically used in combination with ECFP4 and MACCS fingerprints, respectively.²¹

Matched Molecular Pairs

A variety of molecular representations and similarity coefficients have been utilized to assess compound similarity in the SAR analysis. However, for medicinal chemistry applications, the outcome of such whole-molecule similarity calculations is often difficult to chemically reconcile. In general, when different fingerprints are utilized, different similarity values will be obtained.^{1,2} Hence, compounds that are considered similar on the basis of one fingerprint representation might not be classified as similar when other fingerprints are used. Furthermore, as pointed above, no generally applicable similarity thresholds exist.²⁰

To depart from the whole-molecule and global similarity techniques, the concept of matched molecular pairs (MMPs)²² has been introduced that is independent of subjectively determined similarity thresholds and conveys a local molecular similarity perspective. This framework provides a consistent and generally applicable basis to establish structural relationships between compounds, and relate chemical modifications to changes in biological activity. In recent years, this formalism has become increasingly popular and has a significant im-

pact on a number medicinal chemistry applications, especially on the large-scale SAR exploration.

In general terms, an MMP is defined as a pair of compounds that can be interconverted into one another by a well-defined chemical transformation, i.e., the exchange of a substructure. Consequently, molecules forming MMPs are structurally related, yet the structural relationship is not a priori defined. More precisely, the term MMP refers to two compounds that are only distinguished by a small structural modification at a single site, also termed single point MMP. An example of an MMP is depicted in Figure 1.2. Exchanged substructures defining the chemical transformation are highlighted in blue.

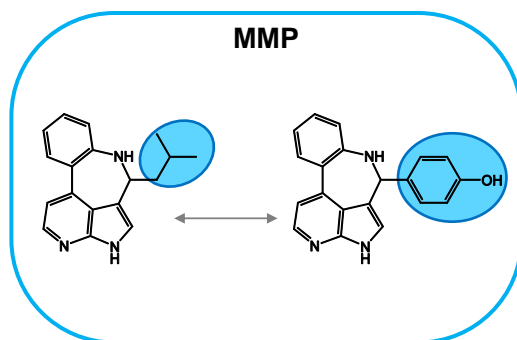


Figure 1.2: Matched molecular pair. Shown are two compounds forming a matched molecular pair (MMP). Exchanged substructures are highlighted.

A large spectrum of algorithms have been developed to systematically generate MMPs. Regardless of methodological details and varying applications, these methods can be categorized into two major classes, i.e., maximum common substructure- (MCS-) based and fragmentation-based methodologies. MCS-based approaches^{23,24} translate the task of finding small structural modifications between a pair of compounds to finding the largest shared substructure between these molecules. It can be accomplished by performing a MCS search (MCSS). Despite many successful applications on large data sets,²³ the MCSS represents a special case of subgraph isomorphism problem that is known to be NP-complete. Furthermore, pairwise compound comparisons are required, which further increases the computational complexity.

Alternatively, fragmentation-based approaches have been introduced to identify shared substructures between a pair of compounds. In general, these approaches can be viewed as a two-step procedure. First, all compounds are subjected to a fragmentation process. Second, by subsequent indexing of the detected substructures, compounds are identified that share a common substructure. Fragmentation-based algorithms are computationally more efficient than MCS-based approaches as each molecule is processed only once.

One of the most widely used fragmentation-based approaches has been introduced by Hussain and Rea.²⁵ Here, molecules are fragmented by systematically deleting all single non-ring bonds (single cuts) between two non-hydrogen atoms, as well as two- and three-bond (double and triple cuts) combinations, resulting in different numbers of fragments. An index table is created to store fragments for each molecule, in which the larger substructure are deposited as keys and the remaining smaller substructures as values. In this way, MMPs can be effectively identified by searching the table for keys with more than one value. To confine the MMPs to only structurally analogous compounds that are only distinguished by a functional group or a single ring system, transformation size-restricted MMPs have been introduced.²⁶

The most prominent feature of the MMP formalism is that it provides a basis for a descriptor-independent, metric-free, and chemically intuitive way to assess structural similarity of bioactive compounds. Hence, it circumvents, at least to some extent, the subjective nature of similarity calculations based on molecular fingerprints.

The exploration of SARs contained in sets of bioactive compounds is a hot spot topic in medicinal chemistry. Yet, the question of *what* represents important SAR information and *how* to best extract and evaluate it is challenging, for several reasons. Molecular representations and structural similarity assessments provide the fundamental basis for SAR analysis. However, depending on the chosen molecular representation and similarity metrics, the outcome of the SAR study may substantially vary. In addition, depending on the size (large sets vs sets of limited size), composition (homogeneous vs. structurally diverse) and origin (HTS vs. compound optimization data) of the data set under investigation, the SAR analysis can, and essentially must, be approached in different

ways. In many instances the outcome of the investigation of SARs is driven by the scientist’s intuition, experience and field of expertise.^{27–30} In the following section, conventional and currently available approaches to explore and exploit SARs contained in data sets will be introduced and discussed.

SAR Exploration

Traditionally, the SAR analysis has been mostly focused on individual compound series, i.e., on structurally homogeneous compounds active against a given target. At late stages of compound design, optimization efforts typically focus only on analogs of a single chemotype. When only a limited number of structurally analogous compounds are available, SARs can be effectively explored on a case-by-case basis. To these ends, R-group tables are utilized that represent the conventional and still most widely used data structure to study the effect of small structural modifications on compound potency (or other properties). On the basis of molecular graphs of the underlying analogs, R-group tables are generated that display the substituents of individual compounds and the corresponding compound activity.

Despite their simplicity, R-group tables become infeasible for structurally heterogeneous compounds or data sets of large size.²⁷ Such tools cannot provide a comprehensive readout of the underlying SARs, and more advanced computational approaches become indispensable for SAR analysis.

Large-Scale SAR Analysis

Since the 1960s, numerous computational methods have been developed to assist in the systematic exploration of SARs contained in a data set. These methods can be roughly classified as *predictive*, i.e., attempting to ultimately predict biological activity, and *descriptive*, i.e., methods that primarily aim to deconvolute and/or visualize SAR information and further identify SAR determinants.

Currently available approaches mostly, but not exclusively, rely on the quantitative SAR paradigm, and hence, are predictive in nature. Powerful

and widely used computational approaches include classical quantitative SAR (QSAR) models.³¹ The ultimate goal of QSAR approaches is the prediction of biological activity for novel, as of yet untested compounds. Using statistical approaches, QSAR methods attempt to establish a (linear) correlation between the biological activity of compounds and their structural or chemical properties. The underlying hypothesis is that if a linear relationship can be derived for a set of known active compounds, then this model can be applied to predict, in quantitative terms, the potency of newly designed analogs. Common to all QSAR methods is that they conceptually rely on one of the fundamental principles in chemoinformatics and medicinal chemistry, the so-called *similarity property principle*² (vide supra).

As a computational technique, QSAR analysis is in principle applicable to (very) large compound data sets. However, this approach is intrinsically limited to structurally homogeneous data sets for which linear relationships can be more reliably derived than for data sets containing more structurally diverse compounds. Hence, test compounds of a different chemotype than the reference molecules fall outside of the applicability domain of most QSAR models, and their activity cannot be reliably predicted.³² Furthermore, it cannot be assumed that SARs are in general linear in nature.

Activity Landscapes

Going beyond QSAR-based predictive methodologies, *activity landscape* models have been developed that systematically combine structural and activity information. These powerful computational models are descriptive in nature and can be used to conceptualize SAR characteristics.

In general terms, activity landscapes can be regarded as any graphical representation that integrates structural and potency similarity relationships between compounds sharing the same biological activity.²⁷ Typically, chemical reference spaces generated from numerical descriptors of molecular structures and other molecular properties serve as a basis for activity landscape models. Each descriptor corresponds to one dimension in the chemical reference

space. Therefore, a set of N descriptors comprises a chemical space of N dimensions. Such high-dimensional space can be further transformed into a human-accessible two-dimensional one with the aid of dimension reduction techniques.^{33,34} Subsequently, bioactive compounds are projected onto the x/y -plane to study the relationship between their molecular properties. The distances between compounds principally relate to the structural similarity of compounds. Hence, structurally similar compounds have shorter distances between them in the space.

In medicinal chemistry and chemoinformatics, activity landscapes are one of the preferred and widely used models to study large-scale SARs. As graphical representations they provide an intuitive access to global and/or local SAR information contained in compound data sets under investigation, and hence facilitate compound selection for further chemical exploration and compound design.

3D Activity Landscapes

Maggiara and colleagues³⁵ envisioned activity landscapes as topological maps that are reminiscent of actual geographical landscapes. These maps represent one of the most prominent types of landscape models, i.e., the hypothetical 3D activity landscapes. Essentially, 3D activity landscapes are generated by adding an activity hypersurface to a set of compounds projected on a 2D chemical reference space. Activity hypersurface provides information about compound potency distribution and compounds with comparable or significantly different potency values can be clearly observed in 3D activity landscapes. Recently, such 3D models have been generated for actual compound sets and their topology has been extensively studied.³⁶

The Nature of SARs

The major goal of SAR exploration is to elucidate how biological activity responds to structural changes. Importantly, different SAR phenotypes can be conceptualized with the aid of idealized 3D activity landscapes and visualized.

Depending on the underlying SAR characteristics of the data set compounds, activity landscapes can be either smooth and easily traversed or may have rugged surfaces.

In general, there are three major SAR categories: continuous, discontinuous, and heterogeneous SARs.¹ Presence of continuous SARs is indicated by gradual changes in compound structures leading to moderate changes in their potency.²⁷ Furthermore, continuous SARs correspond to smooth regions or gently rolling hills in activity landscapes as shown in Figure 1.3a. This type of SARs is consistent with the SPP (*vide supra*). Therefore, continuous SARs provide the conceptual basis for similarity searching and ligand-based virtual screening.³⁷ From a medicinal chemists' point of view, SARs with predictable potency progression are of high interest in compound design.²⁷ In such cases, SAR continuity is an essential consideration.

In contrast, small structural changes resulting in large differences in potency account for discontinuous SARs. The discontinuous character of a set of compounds is represented by rough regions in activity landscape models as illustrated in Figure 1.3b. In hit-to-lead optimization campaigns, SAR discontinuity plays a crucial role, and compounds falling into highly discontinuous regions represent focal points for further chemical exploration.

It is frequently observed that continuous and discontinuous SARs coexist in compound sets sharing the same biological activity.³⁸ Accordingly, the combination of continuity and discontinuity in a single data set is considered to represent heterogeneous SARs. Activity landscapes characterized by heterogeneous SARs are also termed variable activity landscapes (Figure 1.3c).¹ Hence, SAR characteristics of bioactive compounds are essentially continuous, discontinuous, or heterogeneous in nature.^{27,39}

Numerical SAR Analysis

The systematic SAR analysis can also be addressed by introducing numerical functions to quantify the SAR information contained in sets of bioactive compounds. In general terms, the SAR functions are based on pairwise calculations

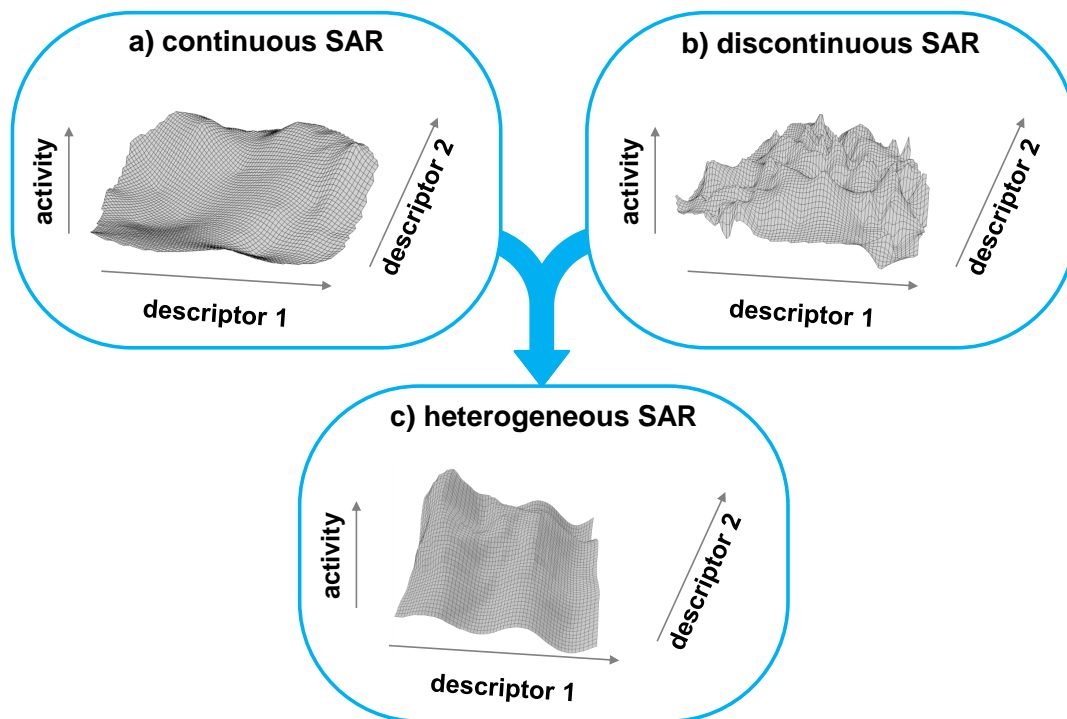


Figure 1.3: SAR phenotypes. Hypothetical 3D activity landscapes of different SAR phenotypes (a, continuous; b, discontinuous; c, heterogeneous) are shown. In these landscapes, compound potency is added as a third coordinate to the 2D projection of the original (high-dimensional) chemical space. Potency distributions are hypothetical. Distances in the 2D projection reflect structural dissimilarity. Adapted from [27].

of structural and activity similarity for data set compounds. Prominent examples include the SAR Index (SARI)⁴⁰ and the Structure-Activity Landscape Index (SALI).⁴¹

The SARI score is calculated for a set of compounds and is a composite of individual SAR continuity and SAR discontinuity scores. A three-step procedure is applied to obtain the final data set score. First, raw scores are calculated, as introduced below. These scores are subsequently transformed into Z-scores on the basis of a panel of reference activity classes. Finally, cumulative probabilities are calculated to map Z-scores onto the value range [0, 1].

For a given data set A , the raw (non-normalized) continuity ($cont_{\text{raw}}(A)$) and discontinuity ($disc_{\text{raw}}(A)$) scores are defined as follows:

$$cont_{\text{raw}}(A) = \frac{\sum_{\{i,j|i>j\}} w_{ij} \frac{1}{1+\text{sim}(i,j)}}{\sum_{\{i,j|i>j\}} w_{ij}}, w_{ij} = \frac{P_i \cdot P_j}{1 + |P_i - P_j|}$$

and

$$disc_{\text{raw}}(A) = \frac{\sum_{\{i,j|\text{sim}(i,j)>T,|P_i-P_j|>1,i>j\}} |P_i - P_j| \cdot \text{sim}(i,j)}{|\{i,j|\text{sim}(i,j) > T, |P_i - P_j| > 1, i > j\}|}$$

where P denotes potency, T a similarity threshold value, and $\text{sim}(i, j)$ the calculated fingerprint similarity for two data set compounds i and j .

The raw continuity score is calculated as the mean of potency weighted pairwise compound dissimilarity and accounts for the presence of structurally dissimilar compounds having high potency, yet small potency differences. On the other hand, the raw discontinuity score is defined as the average of the product of the pairwise potency difference between compounds and their structural similarity. Accordingly, it emphasizes structurally similar compounds having significantly different potency. For discontinuity score calculations, a similarity threshold is selected to limit the calculation to only structurally similar compound pairs. As indicated above, no generally applicable threshold values exist. However, for SAR analysis, a MACCS Tc threshold of 0.85 is typically used to indicate structural similarity.⁵ Furthermore, a potency difference cut off of one is applied to focus on compounds with more than one order of magnitude difference.

The final SARI score is then calculated on the basis of normalized scores and defined as

$$\text{SARI}(A) = \frac{1}{2}(\text{cont}_{\text{norm}}(A) + (1 - \text{disc}_{\text{norm}}(A)))$$

thereby balancing the relative contributions of individual scores. The SARI score ranges between 0 and 1 where high values correspond to predominantly continuous SAR and low values to mainly discontinuous SAR.

Furthermore, the global discontinuous score has been modified to obtain a local, per-compound score. For example, for a given data set compound i , its raw discontinuity score is defined as

$$disc_{\text{raw}}(i) = \frac{\sum_{\{j|i \neq j, \text{sim}(i,j) > T\}} |P_i - P_j| \cdot \text{sim}(i, j)}{|\{j|\text{sim}(i, j) > T, i \neq j\}|}$$

where P denotes potency, T a similarity threshold, and $\text{sim}(i, j)$ the calculated fingerprint similarity between i and its structural neighbors. Similar to its global counterpart, the raw scores are converted into Z-scores by using the intra-set score distribution and then normalized by calculating the cumulative probability on a normal distribution, ultimately mapping the score onto the range $[0, 1]$. Accordingly, it quantifies the contribution of individual compounds to the global data set discontinuity. The score is derived from the average pairwise potency differences of compounds multiplied by their structural similarity. Local structural neighborhoods are typically calculated on the basis of fingerprints. In contrast to the global discontinuity score, all structural neighbors (with respect to a given fingerprint and a similarity threshold value) of a given compound are included in the calculation of its local discontinuity score. Accordingly, a compound obtains a high local discontinuity score if its structural neighbors have significantly different potency values.

Numerical functions provide a quantitative measure of the SAR information content present in sets of bioactive compounds. Raw scores, as introduced above, are normalized with respect to the score distribution in the activity class under investigation. Hence, care must be taken to select a suitable molecular representation as it will inevitably affect, in a characteristic manner, the final data set score.

Therefore, the SARI score can be used as a diagnostic of different SAR phenotypes for activity classes. These functions often complement the landscape-based SAR analysis. Over the past years, activity landscapes have become increasingly attractive tools to assess SAR information contained in compound data sets and gained a lot of interest in the medicinal chemistry and pharmaceutical research. As graphical representations, these models help to view

different SAR information in context and provide intuitive and direct access to SAR characteristics of compound data sets.

Classical Activity Landscape Views

One of the earliest and simplest, and still widely utilized 2D activity landscape representations is the Structure-Activity Similarity (SAS) map.⁴² A prototypical SAS Map is shown in Figure 1.4. In a SAS map, structural similarity of data set compounds is plotted against their activity similarity. Typically, fingerprints are used as molecular representations and the popular Tc⁶ as the similarity metric. A unit data point in the map represents a pair of compounds for which structural and activity similarity relationships are systematically determined. In the schematic representation in Figure 1.4 structural similarity is shown on the x -axis and Tc values vary between 0 (indicating low similarity) and 1 (indicating high similarity). Activity similarity can be represented in different ways, for example, as logarithmic potency difference (e.g., ΔpIC_{50} or ΔpK_i) or normalized potency difference ranging between 0 (identical compound activities) and 1 (maximal potency difference).

A key feature of SAS maps is that it provides the basis for the classification of compounds with different activity landscape features. Selected activity and structural similarity thresholds subdivide the map into four different regions corresponding to four activity landscape features associated with different degrees of SAR information content:

1. Compound pairs at the upper-left region, commonly termed *featureless pairs*, are characterized by low structural and activity similarity. They are not SAR informative and, therefore, of least importance to the SAR analysis.
2. The lower-left region is populated by structurally diverse compound pairs with similar activity. This section corresponds to *similarity cliffs*. From an information-theoretic point of view this is the most prevalent and hence the least informative activity landscape feature. Yet, similarity cliffs can

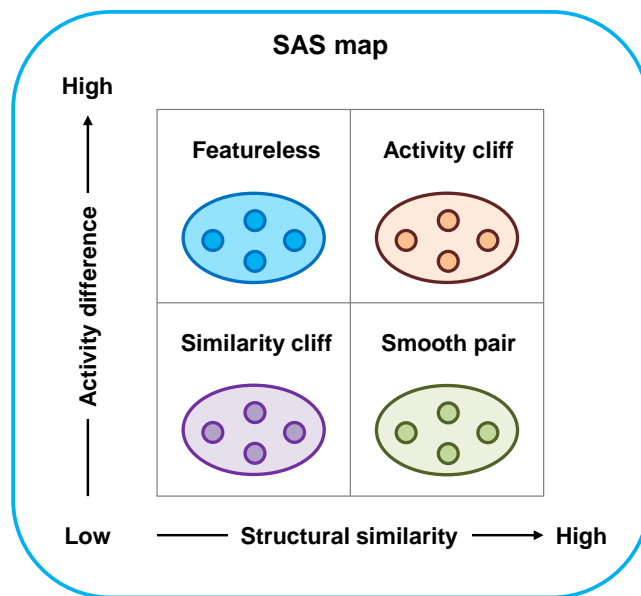


Figure 1.4: Structure-Activity Similarity maps. Shown is a schematic illustration of a Structure-Activity Similarity (SAS) map. On the basis of SAS maps four different activity landscape feature regions are identified. For each compound pair, structural and activity similarity is calculated, thereby uniquely mapping the pair to one of the four regions. Adapted from [43].

aid in the identification of new class of active compounds with similar activity, and are therefore considered SAR informative.

3. The lower-right section consists of structurally related compounds with similar activity. Importantly, these pairs characterize the presence of SAR continuity (small structural modifications lead to moderate changes in activity) and are commonly referred to as *smooth pairs*.
4. *Activity cliffs*^{21,44–46} are generally defined as structurally similar compounds having significant potency difference and populate the upper-right region in the SAS map. As such, they represent an extreme form of SAR discontinuity⁴² and are the most prominent activity landscape feature. Activity cliffs are often rarely present in compound data sets, yet they are focal points of SAR analysis as they directly link structural modification to compound potency improvement.

Hence, activity landscape features are represented by pairs of compounds having varying well-defined structural and activity relationship, typically on the basis of selected structural and activity similarity thresholds. Although they are probably best distinguished on the basis of SAS maps, they are integral part of any activity landscape representation and can also be explored using other models.

SAR Network Modeling

In addition to the simple SAS maps, molecular network representations have been developed to organize and display structural and activity relationships among sets of bioactive compounds, including the Network-like Similarity Graphs (NSGs).⁴⁷ In these graphs, all data set compounds are represented as nodes. To account for structural similarity relationships, edges are drawn between two nodes if the structural similarity between the corresponding compounds exceeds a predefined threshold. In addition, nodes are color-coded according to the compound activity. A continuous color spectrum is applied ranging from green (low activity) over yellow (moderate activity) to red (high activity). Furthermore, nodes are scaled in size with respect to the local per-compound discontinuity score (vide supra). Hence, large nodes correspond to compounds with high discontinuity scores that are predominantly involved in activity cliff formation, and thereby having significant contributions to the global discontinuity of the underlying data set. An exemplary NSG illustrating different information layers is shown in Figure 1.5. It should be noted, that the topological arrangement of individual compounds and clusters of compounds has no chemical meaning. The node positions and the edge lengths are determined by a 2D force-directed graph layout algorithm.⁴⁸

NSGs are landscapes of conceptually different design compared to SAS maps. Here, the focus is on elucidating how local SAR features relate to the global SAR character of the data set. In NSGs, compound subsets (clusters) having different local SAR phenotypes can be easily identified.⁴⁷ For example, clusters of similarly colored and sized nodes highlight regions that are continuous in nature. On the contrary, groups of densely connected compounds that

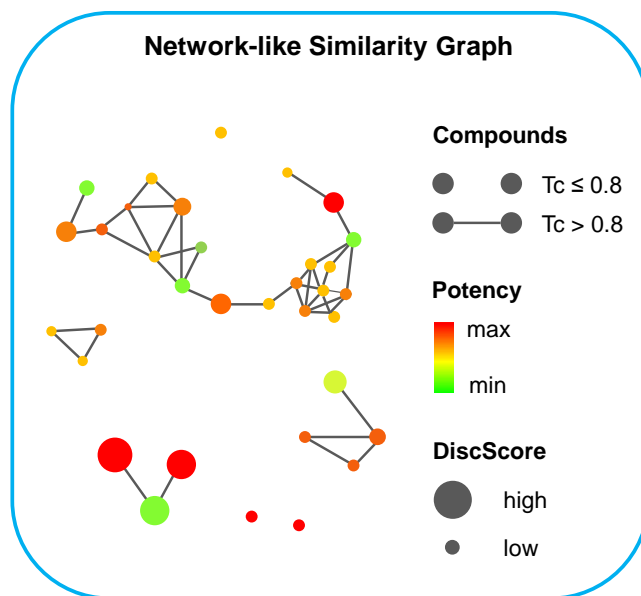


Figure 1.5: Network-like Similarity Graphs. A schematic illustration of a Network-like Similarity Graph (NSG) and its major information layers are shown. Nodes represent compounds and edges reflect structural similarity. In addition, compounds are color-coded according to compound potency (red, high activity; yellow, moderate activity; green, low activity). Furthermore, nodes are scaled to account for local discontinuity scores.

have different colors indicate the presence of local SAR discontinuity. Furthermore, centers of SAR discontinuity can be graphically assessed by selecting large nodes (high local discontinuity score) having many structural neighbors of varying color.

The landscape view provided by NSGs is rich in information layers accounting for different aspects related to global and local SAR characteristics. NSGs are easy to comprehend and navigate, and present one of the most preferred activity landscapes to rationalize SARs in data sets of various size and composition. NSGs have been mostly, but not exclusively, applied to explore compound optimization data. However, the concept has also been successfully applied to high-throughput screens (HTS) typically comprising very large sets of mainly weakly potent compounds.⁴⁹

Approaches to systematic SAR analysis often focus on target-specific compound potency. In lead optimization campaigns, however, potency is only one of several important factors to be considered. A promising drug candidate must

show a desired selectivity profile against a number of targets. An increasing amount of evidence suggests that selective drugs are more the exception rather than the rule and that drugs tend to simultaneously interact with multiple biological targets.⁵⁰ In this respect, selectivity is, in many instances, likely to result from differences in compound potency against multiple targets, rather than from exclusive binding to a single target. For compounds active against multiple targets, the resulting multi-target SARs can be complex and difficult to rationalize. However, they can ultimately reveal a different degree of compound selectivity and hence visualization tools to support a systematic multi-target SAR exploration and exploitation are of high interest. Selectivity NSGs⁵¹ provide a first step towards modification and adaptation of graphical representations to capture structure-selectivity relationships between compounds active against two targets.

Activity Cliffs

Activity landscapes are designed to highlight SAR features and provide graphical access to key compounds for further chemical exploration. Therefore, the study of landscape models and their most prominent feature, the activity cliffs, are central themes in SAR analysis and medicinal chemistry. Activity cliffs,^{21,44-46} as introduced above, are formed by two structurally similar compounds having a large difference in potency. They represent the extreme form of SAR discontinuity and are thought to be rich in SAR information.⁴² Accordingly, their exploration is of prime interest in compound optimization efforts.

In the context of SAR analysis, graphical representations are powerful and indispensable tools, and activity cliffs have mostly been studied using different activity landscape models. The use of activity landscapes greatly benefits from their simplicity, intuitiveness, and the ability to visually prioritize key compounds that predominantly form cliffs.

To gauge the importance and relevance of cliffs in medicinal chemistry, activity cliffs and their distributions have been extensively studied through mining

large databases such as ChEMBL⁵² and Binding DB,⁵³ which represent two major compound data sources for systematic large-scale SAR analysis.

Despite the increasing interest in activity cliff exploration, the definition of activity cliffs is still a matter of debate, for understandable reasons. The two major critical aspects of this formalism are the way the structural similarity is assessed and the notion of “significant difference in potency”. Hence, prior to the systematic assessment of cliffs, structural similarity and activity difference criteria *must* be specified.

Undoubtedly, the most essential task is the assessment of chemical similarity. Tanimoto similarity on the basis of different fingerprint representations has been predominantly used. However, Tanimoto similarity is greatly influenced by the molecular representations used. Hence, different distributions could be obtained when different fingerprints are used as representations.⁴⁶ To circumvent these limitations, attempts have been made to replace the subjective whole-molecule similarity evaluation by more structurally conservative and, from a medicinal chemistry perspective, more chemically intuitive methods. Exemplary substructure-based representations include MMPs (vide supra) and molecular scaffolds.

In addition, data variability also plays an important role in the identification of activity cliffs. Care should be taken to restrict the analysis to only high-confidence data. To these ends, different potency measurements such as equilibrium constants K_i (i.e., theoretically assay-independent) and half maximal inhibitory concentration (IC_{50} ; assay-dependent) should be considered separately. Also, care should be taken when multiple potency measurements are provided for a given compound and a target. In such cases, computing the average, minimum or maximum can be considered as the final potency value for the given compound. However, the choice of final annotation notably affects activity cliff distributions, as it has been previously demonstrated.⁵⁴

Activity Cliff Extensions

The concept of activity cliffs has become increasingly popular in medicinal chemistry. The formalism has been extended in various ways to explore SAR

determinants from different perspectives. Notable extensions are based on different molecular representations that enable assessing the cliff formation at different structural levels. For example, based on the MMP formalism, the notion of MMP-cliffs²⁶ was introduced to limit the cliff analysis to only chemically intuitive and accessible modifications. An MMP-cliff is defined by two compounds that form a transformation size-restricted MMP²⁶ and, in addition, have significantly different potency. Typically 100-fold potency difference (corresponding to 2 orders of magnitude on a logarithmic scale) was considered as a criterion for cliff formation. Transformation-size restricted MMPs ensure that chemical modification distinguishing activity cliff compounds are small.²⁶ An exemplary MMP-cliff is shown on Figure 1.6b.

Recently, activity cliffs have also been defined using molecular scaffolds⁵⁵ (obtained from compounds by removal of R-groups).⁵⁶ On the basis of this categorization, cliffs can be identified having different scaffold/R-group relationships. An R-group based cliff induced by different R-group replacements at the same scaffold is shown in Figure 1.6c.

In general, activity cliffs are explored on a per-target basis. However, it has been frequently observed that many bioactive compounds are active against two or three targets. To these ends, selectivity cliffs⁵¹ were introduced to rationalize dual-target activity (i.e., selectivity) relationships. Precisely, a selectivity cliff is formed by two structurally similar compounds having significantly different activity against their targets. Importantly, this concept represents a first attempt towards multi-target activity cliff exploration. A representative selectivity cliff is shown in Figure 1.6d.

The activity cliff concept has been extended in many different ways.^{26,51,56-58} On the basis of statistical analysis a generally preferred definition has been proposed.⁵⁹ Accordingly, cliff analysis should be confined to only size-restricted MMP-cliffs with potency difference of at least two orders of magnitude. In addition, if available, only equilibrium constants should be considered as potency measurements.

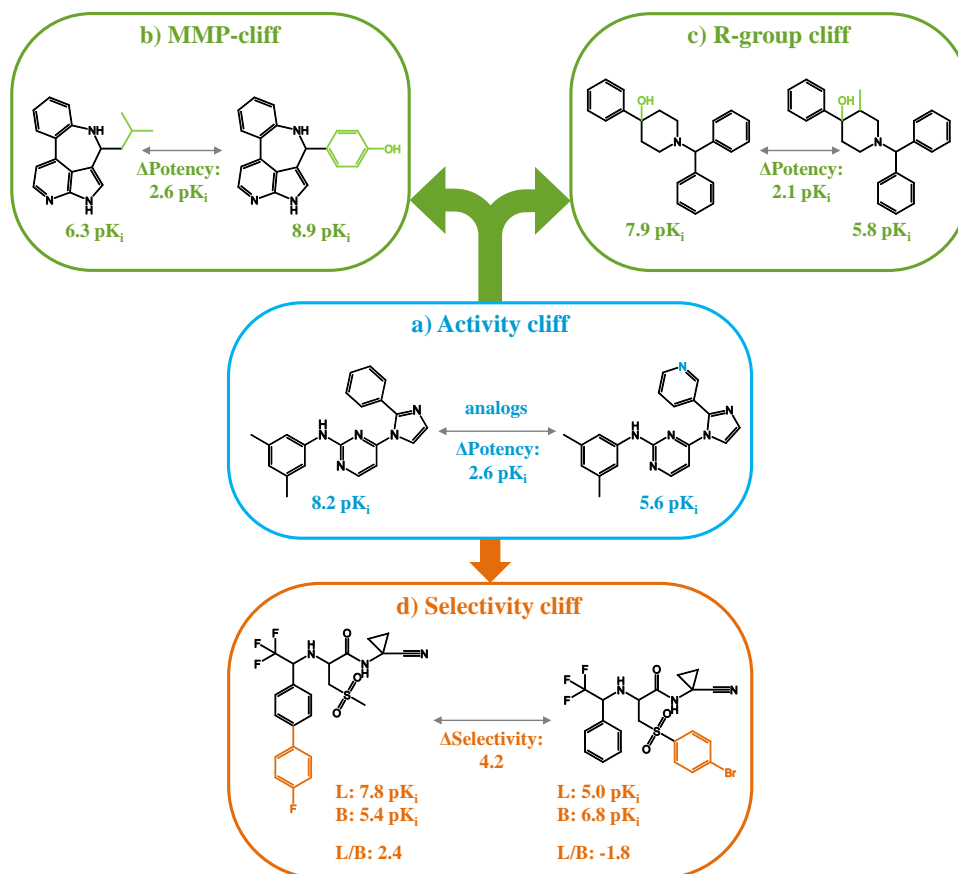


Figure 1.6: Activity cliffs and activity cliff extensions. Shown is an exemplary activity cliff (a) and three representative cliff extensions (b-d). In addition, structural changes between cliff-forming compounds are color-coded and potency values are reported. In (d), compound activity for two targets (cathepsin L and B) are provided. Selectivity scores (L/B) are calculated as the potency difference against the target pair.

Thesis Outline

This work addresses major challenges in systematic SAR exploration in medicinal chemistry and pharmaceutical research. The main focal points have been the design of novel activity landscape models and comprehensive activity cliff analysis.

In this dissertation, eight representative studies are introduced and organized in individual chapters:

- For compounds active against multiple targets, the resulting multi-target SARs are complex and difficult to rationalize. To these ends, a first multi-target activity landscape has been designed to capture multi-target SARs and provide an intuitive graphical access to interesting compounds. The methodology is reported in Chapter 2.
- Activity landscapes introduced in the previous chapter can be meaningfully applied to compounds active against limited number of target (3–5) and are not suitable for compounds with activities against many targets (50 – 100). Chapter 3 introduces the ligand-target differentiation (LTD) map – a first high-dimensional activity landscape model to navigate high-dimensional activity spaces.
- Activity cliffs have been extensively studied, however, it has been unknown how cliffs are distributed in publicly available compounds databases. Chapter 4 investigates the distribution, directionality, and the statistical significance of single- and multi-target activity cliffs formed by currently available bioactive compounds.
- It is well-appreciated that different molecular representations (e.g., molecular fingerprints) inevitably change the numerical assessment of structural similarity, and therefore also the SARs contained in compound data sets. Chapter 5 addresses the influence of representative fingerprints on the SAR information content associated with individual compounds
- Activity cliffs have been studied from many different perspectives. Nevertheless, thus far their utility to aid in the compound optimization efforts has not been systematically analyzed. Chapter 6 introduces the concept of compound pathway models to evaluate the SAR information gain provided by activity cliffs.
- Activity cliffs are thought to contain high SAR information content, thereby providing starting points for further chemical exploration. Chapter 7 addresses, from a chemoinformatics perspective, the relevance and utilization

tion of the activity cliff concept in medicinal chemistry, and its ability to support medicinal chemistry optimization campaigns.

- Recent statistical studies^{46,60} report that the majority of activity cliffs are formed in a coordinated manner and involve multiple active compounds and cliffs. Chapter 8 describes the topology, composition and frequency of occurrence of coordinated cliffs formed by currently available bioactive compounds. Moreover, recurrent topologies are identified and analyzed.
- Chapter 9 reports an extension of the activity cliff concept to capture structure-promiscuity relationships. Furthermore, chemical changes were identified that led to large-magnitude promiscuity effects.

Finally, major findings and key observations of the work presented in this dissertation are summarized and discussed in Chapter 10.

References

- [1] Peltason, L.; Bajorath, J. Systematic Computational Analysis of Structure-Activity Relationships: Concepts, Challenges and Recent Advances. *Future Medicinal Chemistry* **2009**, *1*, 451–466.
- [2] *Concepts and Applications of Molecular Similarity*; Johnson, M., Maggiora, G., Eds.; John Wiley & Sons: New York, 1990.
- [3] Kubinyi, H. Similarity and Dissimilarity: A Medicinal Chemist's View. *Perspectives in Drug Discovery and Design* **1998**, *9–11*, 225–232.
- [4] Eckert, H.; Bajorath, J. Molecular Similarity Analysis in Virtual Screening: Foundations, Limitations and Novel Approaches. *Drug Discovery Today* **2007**, *12*, 225–233.
- [5] Stumpfe, D.; Bajorath, J. Similarity Searching. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1*, 260–282.
- [6] Willett, P. Chemical Similarity Searching. *Journal of Chemical Information and Computer Sciences* **1998**, *38*, 983–996.
- [7] Xue, L.; Bajorath, J. Molecular Descriptors in Chemoinformatics, Computational Combinatorial Chemistry, and Virtual Screening. *Combinatorial Chemistry & High Throughput Screening* **2000**, *3*, 363–372.
- [8] Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31–36.
- [9] Peltason, L.; Bajorath, J. Molecular Similarity Analysis in Virtual Screening. In *Chemoinformatics Approaches to Virtual Screening*, A. Varnek, A. T., Ed.; Royal Society of Chemistry: Cambridge, UK: 2008, pp 120–149.

-
- [10] MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577.
- [11] Ewing, T.; Baber, J. C.; Feher, M. Novel 2D Fingerprints for Ligand-Based Virtual Screening. *Journal of Chemical Information and Modeling* **2006**, *46*, 2423–2431.
- [12] Williams, C. Reverse Fingerprinting, Similarity Searching by Group Fusion and Fingerprint Bit Importance. *Molecular Diversity* **2006**, *10*, 311–332.
- [13] Heikamp, K.; Bajorath, J. Fingerprint Design and Engineering Strategies: Rationalizing and Improving Similarity Search Performance. *Future Medicinal Chemistry* **2012**, *4*, 1945–1959.
- [14] Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naïve Bayesian Classifier. *Journal of Chemical Information and Computer Sciences* **2004**, *44*, 170–178.
- [15] Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *Journal of Chemical Information and Computer Sciences* **2004**, *44*, 1708–1718.
- [16] Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50*, 742–754.
- [17] *Molecular Operating Environment (MOE 2012.10)*; Chemical Computing Group: Montreal, Canada, 2012.
- [18] Jaccard, P. Nouvelles Recherches sur la Distribution Florale. *Bulletin de la Société Vaudoise* **1908**, *44*, 223–270.
- [19] Flower, D. R. On the Properties of Bit String-Based Measures of Chemical Similarity. *Journal of Chemical Information and Computer Sciences* **1998**, *38*, 379–386.
- [20] Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry. *Journal of Medicinal Chemistry* **2013**.

REFERENCES

- [21] Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *Journal of Medicinal Chemistry* **2012**, *55*, 2932–2942.
- [22] Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Chemoinformatics in Drug Discovery*; Wiley-VCH: 2005, pp 271–285.
- [23] Raymond, J. W.; Watson, I. A.; Mahoui, A. Rationalizing Lead Optimization by Associating Quantitative Relevance with Molecular Structure Modification. *Journal of Chemical Information and Modeling* **2009**, *49*, 1952–1962.
- [24] Warner, D. J.; Griffen, E. J.; St-Gallay, S. A. WizePairZ: A Novel Algorithm to Identify, Encode, and Exploit Matched Molecular Pairs with Unspecified Cores in Medicinal Chemistry. *Journal of Chemical Information and Modeling* **2010**, *50*, 1350–1357.
- [25] 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.
- [26] 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.
- [27] Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity Landscape Representations for Structure-Activity Relationship Analysis. *Journal of Medicinal Chemistry* **2010**, *53*, 8209–8223.
- [28] Takaoka, Y.; Endo, Y.; Yamanobe, S.; Kakinuma, H.; Okubo, T.; Shimazaki, Y.; Ota, T.; Sumiya, S.; Yoshikawa, K. Development of a Method for Evaluating Drug-Likeness and Ease of Synthesis Using a Data Set in Which Compounds are Assigned Scores Based on Chemists' Intuition. *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 1269–1275.
- [29] Lajiness, M. S.; Maggiora, G. M.; Shanmugasundaram, V. Assessment of the Consistency of Medicinal Chemists in Reviewing Sets of Compounds. *Journal of Medicinal Chemistry* **2004**, *47*, 4891–4896.

- [30] Kutchukian, P. S.; Vasilyeva, N. Y.; Xu, J.; Lindvall, M. K.; Dillon, M. P.; Glick, M.; Coley, J. D.; Brooijmans, N. Inside the Mind of a Medicinal Chemist: The Role of Human Bias in Compound Prioritization during Drug Discovery. *PLoS one* **2012**, *7*, e48476.
- [31] Esposito, E. X.; Hopfinger, A. J.; Madura, J. D. Methods for Applying the Quantitative Structure-Activity Relationship Paradigm. In *Chemoinformatics – Concepts, Methods, and Tools for Drug Discovery*, Bajorath, J., Ed.; Humana Press: Totowa, NJ, 2004, pp 131–213.
- [32] Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G.; Niemela, J.; Mekenyan, O. A Stepwise Approach for Defining the Applicability Domain of SAR and QSAR Models. *Journal of Chemical Information and Modeling* **2005**, *45*, 839–849.
- [33] Agrafiotis, D. K.; Lobanov, V. S. Nonlinear Mapping Networks. *Journal of Chemical Information and Computer Sciences* **2000**, *40*, 1356–1362.
- [34] Wawer, M.; Lounkine, E.; Wassermann, A. M.; Bajorath, J. Data Structures and Computational Tools for the Extraction of SAR Information from Large Compound Sets. *Drug Discovery Today* **2010**, *15*, 630–639.
- [35] Maggiora, G. M.; Shanmugasundaram, V.; Lajiness, M. S.; Doman, T. N.; Schulz, M.; Oprea, T. *A Practical Strategy for Directed Compound Acquisition*; Oprea, T., Ed.; Wiley-VCH: 2005.
- [36] 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.
- [37] Stumpfe, D.; Bajorath, J. Applied Virtual Screening: Strategies, Recommendations, and Caveats. *Virtual Screening: Principles, Challenges, and Practical Guidelines* **2011**, 291–318.
- [38] Peltason, L.; Bajorath, J. Molecular Similarity Analysis Uncovers Heterogeneous Structure-Activity Relationships and Variable Activity Landscapes. *Chemistry and Biology* **2007**, *14*, 489–497.

REFERENCES

- [39] Stumpfe, D.; Bajorath, J. Methods for SAR Visualization. *RSC Advances* **2012**, *2*, 369–378.
- [40] Peltason, L.; Bajorath, J. SAR Index: Quantifying the Nature of Structure-Activity Relationships. *Journal of Medicinal Chemistry* **2007**, *50*, 5571–5578.
- [41] Guha, R.; Van Drie, J. H. Structure-Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *Journal of Chemical Information and Modeling* **2008**, *48*, 646–658.
- [42] Shanmugasundaram, V.; Maggiora, G. M. Characterizing Property and Activity Landscapes Using an Information-Theoretic Approach. In *Proceedings of 222nd American Chemical Society National Meeting, Division of Chemical Information, Chicago, IL, August 26–30, 2001*; American Chemical Society: Washington, DC, 2001, abstract no. 77.
- [43] Medina-Franco, J. L. Scanning Structure-Activity Relationships with Structure-Activity Similarity and Related Maps: From Consensus Activity Cliffs to Selectivity Switches. *Journal of Chemical Information and Modeling* **2012**, *52*, 2485–2493.
- [44] 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.
- [45] Maggiora, G. M. On Outliers and Activity Cliffs why QSAR Often Disappoints. *Journal of Chemical Information and Modeling* **2006**, *46*, 1535–1535.
- [46] Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. *Journal of Medicinal Chemistry* **2014**, *57*, 18–28.
- [47] 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.

-
- [48] Fruchterman, T. M.; Reingold, E. M. Graph Drawing by Force-Directed Placement. *Software: Practice and Experience* **1991**, *21*, 1129–1164.
- [49] Wawer, M.; Bajorath, J. Extracting SAR Information From a Large Collection of Anti-Malarial Screening Hits by NSG-SPT Analysis. *ACS Medicinal Chemistry Letters* **2011**, *2*, 201–206.
- [50] Jalencas, X.; Mestres, J. On the Origins of Drug Polypharmacology. *Medicinal Chemistry Communications* **2013**, *4*, 80–87.
- [51] Peltason, L.; Hu, Y.; Bajorath, J. From Structure-Activity to Structure-Selectivity Relationships: Quantitative Assessment, Selectivity Cliffs, and Key Compounds. *ChemMedChem* **2009**, *4*, 1864–1873.
- [52] 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.
- [53] Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities. *Nucleic Acids Research* **2007**, *35*, D198–D201.
- [54] 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.
- [55] Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of Medicinal Chemistry* **1996**, *39*, 2887–2893.
- [56] 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.
- [57] 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.

REFERENCES

- [58] Iyer, P.; Stumpfe, D.; Bajorath, J. Molecular Mechanism-Based Network-like Similarity Graphs Reveal Relationships between Different Types of Receptor Ligands and Structural Changes that Determine Agonistic, Inverse-Agonistic, and Antagonistic Effects. *Journal of Chemical Information and Modeling* **2011**, *51*, 1281–1286.
- [59] 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.
- [60] 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.

Chapter 2

Design of Multi-Target Activity Landscapes That Capture Hierarchical Activity Cliff Distributions

Introduction

Understanding SAR characteristics of bioactive compounds is a central task in medicinal chemistry and pharmaceutical research. To facilitate SAR analysis, different activity landscape models have been developed. Regardless of their methodological differences, these methods focus only on a single or at most two targets, in the latter case giving rise to selectivity landscapes. The design of landscape representations for compounds active against multiple targets is a challenging and, as of yet, unsolved task. In this work, a first multi-target activity landscape approach is introduced that is based on a numerical encoding scheme of activity profiles. The model facilitates the identification and selection of compounds, or groups of compounds, involved in multi-target activity cliffs. Furthermore, the contribution of individual compounds to global multi-target SARs can be monitored.

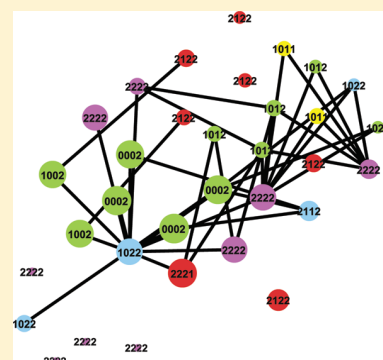
Design of Multitarget Activity Landscapes That Capture Hierarchical Activity Cliff Distributions

Dilyana Dimova, Mathias Wawer, Anne Mai Wassermann, and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit, Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113 Bonn, Germany

S Supporting Information

ABSTRACT: An activity landscape model of a compound data set can be rationalized as a graphical representation that integrates molecular similarity and potency relationships. Activity landscape representations of different design are utilized to aid in the analysis of structure–activity relationships and the selection of informative compounds. Activity landscape models reported thus far focus on a single target (i.e., a single biological activity) or at most two targets, giving rise to selectivity landscapes. For compounds active against more than two targets, landscapes representing multitarget activities are difficult to conceptualize and have not yet been reported. Herein, we present a first activity landscape design that integrates compound potency relationships across multiple targets in a formally consistent manner. These multitarget activity landscapes are based on a general activity cliff classification scheme and are visualized in graph representations, where activity cliffs are represented as edges. Furthermore, the contributions of individual compounds to structure–activity relationship discontinuity across multiple targets are monitored. The methodology has been applied to derive multitarget activity landscapes for compound data sets active against different target families. The resulting landscapes identify single-, dual-, and triple-target activity cliffs and reveal the presence of hierarchical cliff distributions. From these multitarget activity landscapes, compounds forming complex activity cliffs can be readily selected.



INTRODUCTION

The concept of activity landscapes provides the basis for a comprehensive analysis of structure–activity relationships contained in large compound sets.¹ For example, activity landscape models aid in the rationalization of global and local SAR features and the selection of compounds for chemical exploration.¹ Activity landscapes are generally defined as representations that integrate structure and potency relationships between compounds having the same biological activity.² As such, activity landscapes can be represented in rather different ways, ranging from simple 2-D plots that compare the structural and activity similarity between data set compounds in a pairwise manner³ and potency-annotated molecular network representations^{4,5} to detailed 3-D landscape views.⁶ In such 3-D activity landscape models, an interpolated potency surface is added to a 2-D projection of chemical reference space as the third dimension,⁶ giving rise to landscapes that are reminiscent of topographical maps.^{2,7}

Regardless of the specifics of different activity landscape representations, the assessment of pairwise molecular similarity relationships is a key element of landscape design. It has been shown that chosen molecular representations for similarity evaluation very often influence the topology of landscape models.^{6,8} The most prominent features of activity landscapes, however they might be represented, are activity cliffs that are formed by pairs or groups of structurally similar compounds, for

example, analog series, with large differences in potency.^{2,9} Regions spanning multiple activity cliffs are rich in SAR information content and represent primary focal points of landscape analysis.

Although a common feature of most activity landscape representations reported thus far is that they focus on activity against a single target, there are no principal reasons to limit activity landscape modeling to individual targets. However, only very few studies have considered two biological activities of compounds for the generation of activity landscapes. Recently, an extension of the activity landscape concept has been introduced, where potency ratios for compounds active against two targets have been utilized instead of single-target compound potency values.¹⁰ The use of potency ratios (or logarithmic potency differences) to annotate similarity-based compound networks is straightforward, giving rise to selectivity landscapes and the notion of selectivity cliffs that are formed by similar compounds having significantly different potency against the two targets.¹⁰ Comparisons of dual-target activity landscapes have also been carried out in a pairwise manner for analog series with activity annotations against three targets (i.e., yielding three pairwise landscape representations).¹¹ In this case, insights into complex SARs of compound series could be obtained by

Received: December 6, 2010

Published: January 28, 2011

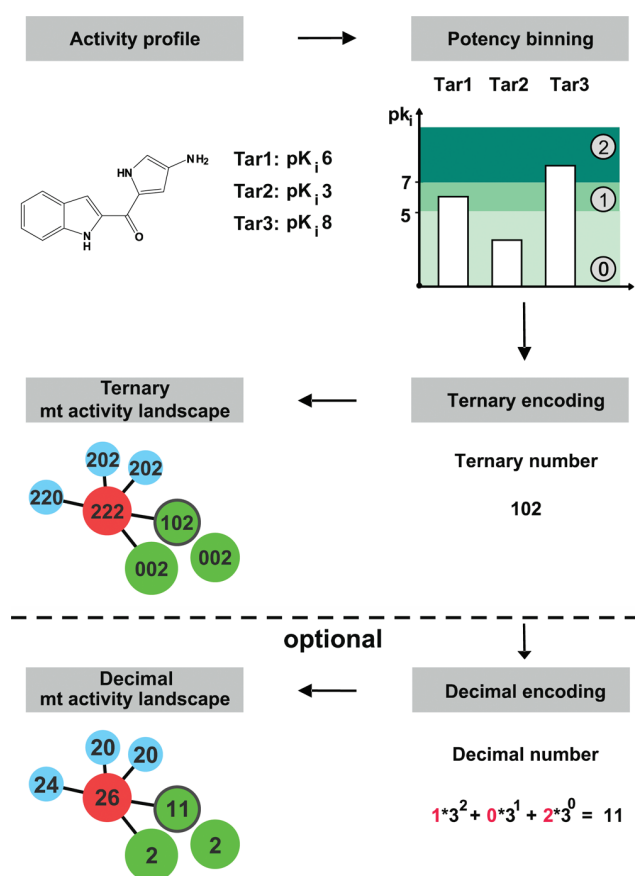


Figure 1. Generation and encoding of compound activity profiles. The schematic illustration summarizes the steps involved in converting compound potencies against multiple targets into activity profiles and representing these activity profiles as ternary numbers (or corresponding optional decimal codes). Ternary (or decimal) codes are used as node labels in the multitarget activity landscape representations.

comparing corresponding regions in target-pair landscape models.¹¹

Going beyond target-pair landscapes and selectivity cliffs, a currently unsolved problem is the generation of an activity landscape framework for multiple targets. Generating multitarget activity landscapes would be relevant, for example, for the study of ligands active against different members of protein families or polypharmacological targets. However, activity landscape representations for compounds active against three or more targets cannot be obtained on the basis of currently available models, and new design concepts are required. Herein, we report a methodology to derive and visualize multitarget activity landscapes and analyze activity cliff distributions. As an exemplary application, the approach is utilized to characterize compound data sets active against members of different target families.

MATERIALS AND METHODS

Potency Binning and Encoding. A three-level scheme is applied for encoding compound activity profiles.

- (1) Potency values are assigned to three different ranges (bins) in order to classify compounds as *weakly potent* ($pK_i \leq 5$), *moderately potent* ($pK_i > 5$ and $pK_i \leq 7$), or *highly potent* ($pK_i > 7$). Hence, a compound is considered

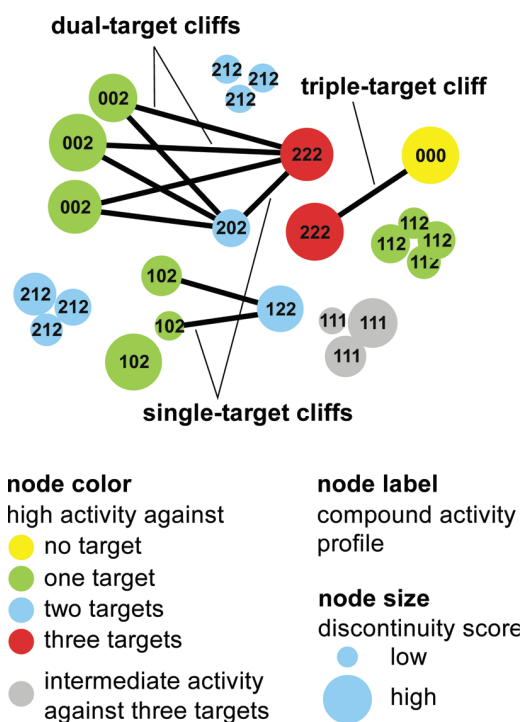


Figure 2. Activity landscape representation. The design of multitarget activity landscapes and the components and information layers of these network-like graphs are schematically illustrated.

weakly potent against a given target if its potency is less than or equal to $10 \mu M$, moderately potent if its potency is greater than $10 \mu M$ but lower than or equal to 100 nM , and highly potent if its potency is higher than 100 nM (these potency intervals can be adjusted for different applications).

- (2) Weakly potent compounds are assigned the ternary digit “0”, moderately potent the digit “1”, and highly potent the digit “2”. These potency bin values define a ternary numeral system. A ternary numeral system represents numeric values using only the digits 0, 1, and 2. To simplify the notation, we denote a ternary number v of a length n by a sequence of ternary digits of length n , i.e., $v = v_1 \dots v_n$ with $v_i \in \{0, 1, 2\}$ for all $1 \leq i \leq n$.
- (3) The activity profile of a compound active against n targets is uniquely mapped to a ternary number of length n . For a given ternary number v and number of targets n , we denote by $[v]_i$ the ternary digit at position i in v . This ternary number can also be converted into a decimal code that serves as a compound label defining its multitarget activity profile.

Figure 1 illustrates our three-level potency classification and encoding scheme.

Multitarget Graphs. For graphical representations of multitarget activity landscapes, we have modified and further extended the network-like similarity graph (NSG) data structure,⁵ a JAVA implementation that is publicly available as part of the SARANEA program suite.¹² For multitarget landscape displays, this similarity-based compound network was annotated with compound activity profiles. For visualization, a layout algorithm¹³ is applied that places groups of densely connected vertices in close vicinity, while separating weakly connected regions of the graph from

Table 1. Compound Data Sets^a

set	activity	no. of compounds	no. of targets	targets
AR	adenosine receptor antagonists	342	3	adenosine receptors A1, A2a, A3
MT	monoamine transporter inhibitors	299	3	dopamine (DA), norepinephrine (NE), serotonin (5HT) transporters
OR	opioid receptor antagonists	98	4	δ -, κ -, μ -opioid receptor, nociceptin (O) receptor
CA	carbonic anhydrase (CA) inhibitors	96	4	carbonic anhydrases 1, 2, 9, 12

^a Four compound data sets were collected from ChEMBL. For each set, the specific activity (activity), number of antagonists or inhibitors (no. of compounds), number of targets (no. of targets), and target names (targets) are reported.

each other. Multitarget graphs are interactively navigated. They can be zoomed and edited, and nodes are graphically associated with compound structures (a structure is displayed when the cursor is placed on a node). The data structure can also be systematically searched for structural relationships or activity cliffs.

Figure 2 summarizes the design elements of these multitarget graphs. Compounds are represented as nodes that are connected by type-1 edges if their pairwise structural similarity, calculated as Tanimoto similarity for ECFP4 fingerprints,¹⁴ exceeds a threshold of 0.4. Nodes are labeled with ternary (or decimal) codes that represent their activity profiles and are color-coded to highlight selected profiles: compounds highly potent against only one, two, three, or four targets are colored green, blue, red, and purple, respectively, and compounds with intermediate potency against all targets are colored light gray. All compounds with potency profiles different from these four categories (i.e., different combinations of moderate and/or low potencies) are colored yellow. Nodes are scaled in size according to a multitarget discontinuity score (as described below) such that large nodes represent high discontinuity score values. Single-target and multitarget activity cliffs formed by pairs of compounds are identified by type-2 edges, i.e., an edge is drawn between two compounds if they form an activity cliff. Single-target and multiple target activity cliffs are selectively displayed.

Multitarget Discontinuity Score. A multitarget discontinuity score (mtDiscScore) is defined to quantitatively account for the degree of multitarget SAR discontinuity that an individual compound introduces in the activity landscape. This score is a variant of the SAR Index per-compound discontinuity score we previously introduced.⁵ For conventional single-target activity landscapes, the per-compound discontinuity scores identifies compounds that have large potency deviations from their immediate structural neighbors. Here, the mtDiscScore is defined in analogy to the per-compound discontinuity score. Thus, the mtDiscScore quantitatively compares potency differences across multiple targets for each data set compound with its structural neighbors in a pairwise manner. Formally, let n be the number of targets. Two compounds are considered similar if their ECFP4 value is greater than 0.4. For every compound c we define the set $N(c)$ as the set of all structural neighbors. Furthermore, let $|N(c)| = m$. For a ligand c , we define $[c]_i$ as the potency value of c against target i . Then the mtDiscScore is defined as

$$\text{mtDiscScore}_{\text{raw}}(c) = \frac{1}{n \cdot m} \sum_{c' \in N(c) \text{ sim}(c, c') > 0.4} \sum_{i=1}^n |[c]_i - [c']_i| \cdot \text{sim}(c, c')$$

The raw scores are standardized on the basis of Z-score calculations including all compounds in a data set and normalized

by calculating the cumulative probability for a normal distribution, yielding final scores falling into the range [0,1].

Given this formalism, a compound makes large contributions to multitarget SAR discontinuity (and achieves a high score) if it has many structural neighbors with different potency profiles.

Data Sets. For the introduction of multitarget activity landscapes, we selected four compound data sets from ChEMBL¹⁵ with (antagonistic or inhibitory) activity against three or four targets belonging to four different families, i.e., adenosine receptors (AR), monoamine transporters (MT), opioid receptors (OR), and carbonic anhydrases (CA). The composition of these compound sets is summarized in Table 1. Only potency measurements were selected with the highest target confidence level (i.e., target confidence score 9) for direct interactions (i.e., target relationship type "D"). Potency measurements containing threshold values (i.e., reported as > or <) were not considered. For compounds with multiple potency values reported against the same target, the arithmetic mean was calculated to yield the final potency.

RESULTS AND DISCUSSION

Activity Landscape Design Principles and Applications.

The generation of activity landscape models generally requires the integration of compound similarity and potency relationships. Assessing similarity relationships is independent of the number of targets and a constant for different landscape models representing the same data set. Conventional single-target activity landscapes are utilized to extract SAR information from compound data sets and identify prominent activity cliffs, i.e., structurally similar compounds having large potency differences. The major challenge of multitarget activity landscape design is how to combine compound potency relationships for several targets and best represent activity profiles. For two targets, this can be accomplished by assigning potency ratios to compounds instead of individual potency values. However, for more than two targets, this simple approach is no longer feasible and different data structures are required. For this purpose, we introduce herein a multitarget potency encoding scheme that is based on a ternary numeral system.

The representation of multitarget activity landscapes does not only provide a methodological challenge, but is also of practical relevance for pharmaceutical research. For example, in order to selectively optimize compounds against an individual target compared to closely related ones, multitarget landscapes provide an access to multitarget activity cliffs, their most prominent features. A multitarget activity cliff is formed by compounds with differential potency against two or more targets. Compounds forming such cliffs are most likely to provide information about structural modifications that change compound potency against two or more targets in either the same or opposite directions. This information would be valuable to guide selective

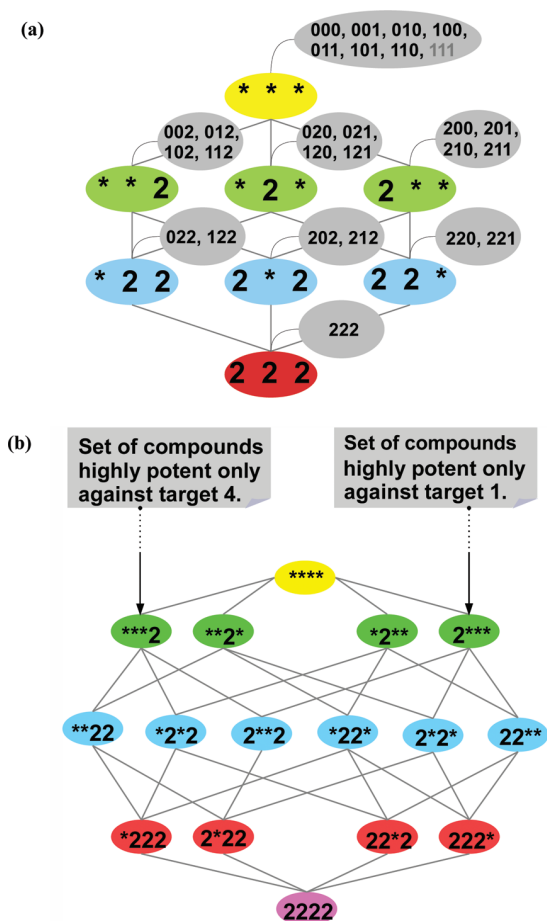


Figure 3. Activity profiles. Shown is the formal organization of activity profiles containing highly potent compounds for (a) three and (b) four targets. Asterisks indicate ternary digits of either 0 or 1. In (a), sets of all ternary numbers covered by the generic profiles are shown as gray tags. The colors correspond to the node coloring scheme introduced for multitarget activity landscapes. This organization scheme provides a basis for the systematic enumeration of all principally possible single-target and multitarget activity cliffs and specification of different activity cliff types using decimal code combinations.

optimization efforts. In multitarget landscapes, as introduced herein, compounds forming such cliffs are readily identified, and their potency profiles can then be directly compared.

Activity Profile Encoding. By classifying compounds as weakly (“0”), moderately (“1”), or highly potent (“2”), we encode multitarget activity profiles as ternary numbers, e.g., “121” for three or “0112” for four targets. Applying this simple formalism, each possible compound activity profile is uniquely encoded. The activity profiles can be systematically organized in different ways. For example, in Figure 3, activity profiles are arranged on the basis of high compound potency against one to three (Figure 3a) or one to four (Figure 3b) targets, which is a prerequisite for the systematic organization of all theoretically possible activity cliffs, as discussed in the following section. The top level in these graphs represents activity profiles of compounds that do not have high potency against any target, the second level compounds with high potency against only one target, the third level compounds with high potency against two targets, and so on.

Activity Cliff Organization. On the basis of our encoding scheme, we define that an activity cliff is formed by any pair of

compounds representing a “2–0” potency combination against any target, i.e., a cliff is formed by one compound with high and one with low potency. Hence, for compounds with activity against three targets, we can formally distinguish between single-, dual-, and triple-target activity cliffs, and for compounds active against four targets, quadruple-target activity cliffs can in principle also be formed. In Figure 3a, all possible activity profiles for highly potent compounds and three or four targets are reported that can participate in the formation of activity cliffs. For three and four targets, there are a total of 7 and 15 high-potency profile categories, respectively, that can form activity cliffs. In Figure 3a, the ternary codes of activity profiles representing each type are also provided. These activity profiles can be systematically paired with “0”-containing profiles to yield the theoretically possible numbers of single- and multiple-target activity cliffs. It should be noted that compounds with moderate potency against all targets cannot participate in the formation of activity cliffs. The corresponding activity profiles for three and four targets are “111” and “1111”, respectively.

We have calculated the numbers of all principally possible single- and multiple-target activity cliffs that result from unique profile combinations. For three targets, there are 147, 42, and 4 different types of single-, dual-, and triple-target activity cliffs possible, respectively. For four targets, the corresponding numbers are 1372 single-, 588 dual-, and 112 triple-target cliffs. In addition, in this case, 16 types of quadruple-target cliffs could be formed. Each of these potential activity cliffs is identified by a unique code combination. For example, activity profile “222” (representing compounds with high potency against three targets) forms a single-target activity cliff with profile “201”.

For three or four targets, ternary numbers might be directly used as node labels. If more target annotations would be available, ternary codes might become too large, and hence, they could be transformed into shorter decimal codes for visualization (as illustrated in Figure 1), for example, using “16” instead of “121” or “14” for “0112”. Decimal codes are less intuitive than ternary numbers, but can be interpreted, for example, with the help of a conversion table.

Multitarget Activity Landscapes. In Figure 4, the multitarget graph for the MT compound set is shown. For the description and interpretation of our multitarget activity landscape design, we focus on the MT compound set in the text and provide corresponding representations for the three other data sets in the Supporting Information. Figure S1 of the Supporting Information shows the multitarget graphs for the AR, OR, and CA sets. In addition, for the four-target CA set, a graph with decimal node labels is also shown for comparison to illustrate the decimal coding scheme. Detailed descriptions of the graphs of these three data sets are provided in the Results section of the Supporting Information. In all graph representations, type-1 edges indicating pairwise compound similarity relationships are not displayed for clarity.

The MT inhibitor landscape displays extensive clustering of compound subsets, revealing regions with densely packed nodes that are separated from other clusters. In many cases, compounds in individual clusters have similar activity profiles, i.e., they have the same node color and the same or similar codes. The presence of similar activity profiles would be expected for compounds that are active against closely related targets. Several clusters of structurally similar compounds are observed that include predominantly highly potent (red; code 222) or weakly potent (yellow; code 000) compounds. Furthermore, the graph contains

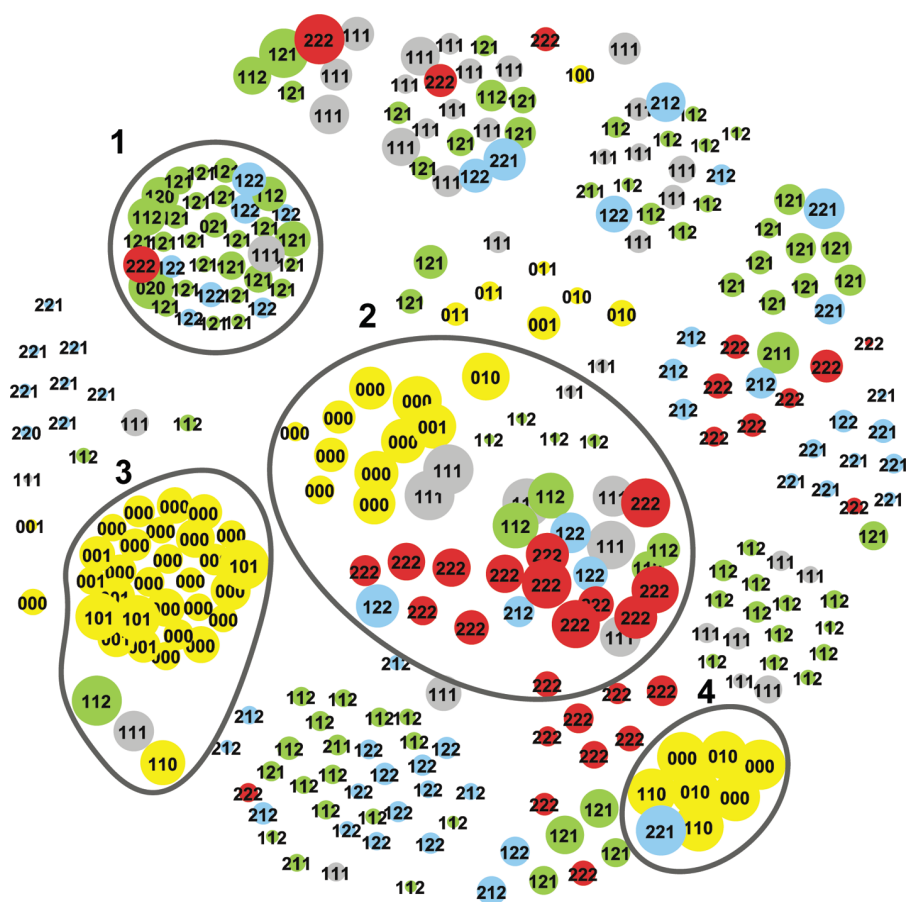


Figure 4. Multitarget graph. Shown is the multitarget activity landscape representation for the MT compound data set. Selected clusters are encircled and shown in detail in Figure 5.

Table 2. Activity Cliff Statistics^a

set	stc	dtc	ttc	activity cliff distribution		
				type	count	degree
AR	121	9	0	212–002	84	single
				222–102	15	single
				222–002	9	dual
MT	32	5	0	112–000	9	single
				122–120	7	single
				122–020	7	single
OR	54	4	0	2112–0001	26	single
				2112–0011	19	single
				2122–0001	4	dual
CA	49	16	15	2222–1012	14	single
				2222–0002	11	triple
				2122–0002	8	dual

^a Four each compound data set, the number of single- (stc), dual- (dtc), and triple-target activity cliffs (ttc) is reported. In addition, “activity cliff distribution” reports the three most frequently occurring activity cliff types for each data set, and “degree” identifies single-, dual-, or triple-target cliffs.

numerous gray nodes (code 111), representing compounds of consistently moderate potency against multiple targets that cannot form activity cliffs according to our classification scheme.

However, the MT activity landscape also reveals clusters that are rich in differently colored nodes representing compounds with different activity profiles. Selected clusters characterized by the presence of rather different activity profiles are encircled in Figure 4. Such regions of heterogeneous node composition are prime candidates for activity cliff formation. Furthermore, the graph contains many differently sized nodes that indicate different compound contributions to multitarget SAR discontinuity, as discussed below.

Activity Cliff Distribution. We next identified activity cliffs contained in all four data sets. The activity cliff distributions are reported in Table 2. In the AR, MT, and OR sets, single- and dual-target cliffs were detected, and in the CA set, single-, dual-, and triple-target cliffs were identified. The activity cliffs were unevenly distributed. For each data set, they involved different target combinations, and not all targets were involved in the formation of cliffs. For AR, MT, OR, and CA, 121, 32, 54, and 49 single-target cliffs were found and 9, 5, 4, and 16 dual-target cliffs, respectively. In addition, for CA, 15 triple-target cliffs were identified. Thus, multitarget activity cliffs were more sparsely distributed than single-target cliffs in these compounds sets directed at closely related members of different protein families. We also found that certain activity profiles were more frequently involved in activity cliff formation than others, and consequently, some activity cliff types were preferentially formed. Table 2 reports the three most frequently occurring activity cliff types for

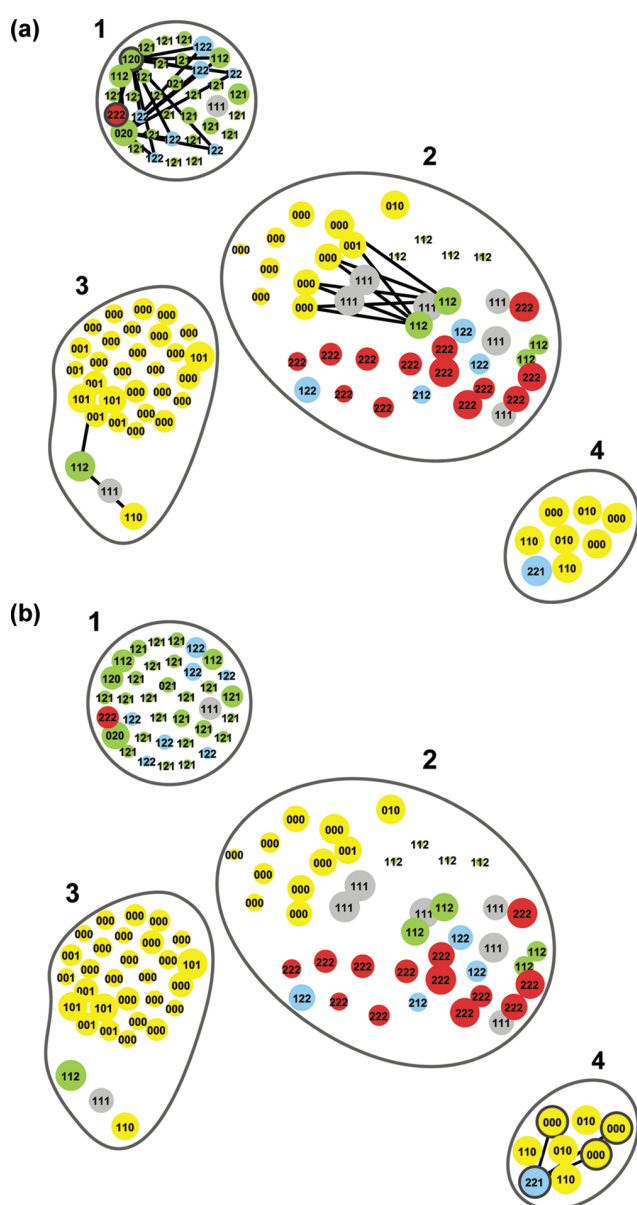


Figure 5. Compound clusters forming activity cliffs. For the MT data set, compound clusters are displayed where prominent single- or dual-target cliffs are formed. In (a), single-target cliffs are reported for the dopamine transporter, and in (b), dual-target cliffs are reported for the norepinephrine and serotonin transporter. Selected activity cliffs are highlighted, and the structures of the corresponding compounds and further details are shown in Figure 6.

each data set. Certain activity cliff types occurred with higher frequency than others, and frequent activity cliffs varied in a data set-specific manner.

Activity Cliff Patterns. We then identified compound clusters in the different data sets where activity cliffs mostly occurred. Figure 5 shows enlarged clusters from the multitarget graph of the MT data set (encircled in Figure 4) that contain single- (Figure 5a) and dual-target cliffs (Figure 5b). In addition, Figure S2 of the Supporting Information shows activity cliff-containing clusters for the other compound sets. In these representations, type-2 edges are displayed, each of which marks an activity cliff. For each data set, single-target activity cliffs involving an exemplary target are

shown, and activity cliff views involving the remaining targets are provided in Figure S3 of the Supporting Information. For the AR, OR, and CA sets, detailed descriptions are given in the Results section of the Supporting Information. We generally observe that compounds have very different node sizes according to multitarget discontinuity scoring. In conventional single-target landscapes, compounds with large nodes typically have a potency that significantly differs from their structural neighbors, introduce local SAR discontinuity, and are most frequently involved in the formation of activity cliffs.⁵ However, in multitarget landscapes, SAR discontinuity is a much more complex phenomenon because of the many different potency relationships that can result from comparisons of multitarget activity profiles. A characteristic feature of the activity cliff distributions in Figure 5 and Figure S2 of the Supporting Information is that compounds with large nodes (i.e., compounds that introduce notable multitarget SAR discontinuity) are often not involved in the formation of well-defined activity cliffs, but that cliffs are also formed by compounds having rather different node sizes. Thus, in multitarget activity landscapes, the introduction of SAR discontinuity and the formation of large-magnitude activity cliffs do not necessarily correlate because of the complexity of potency relationships between compounds active against multiple targets. However, similar to single-target landscapes, key compounds also emerge that make large contributions to multitarget SAR discontinuity and form multiple activity cliffs. For example, in the most densely connected region of the CA data set, the only compound set where triple-target cliffs were detected, overlapping yet distinct subsets of nodes form dual- (Figure S2f of the Supporting Information) and triple-target activity cliffs (Figure S2g of the Supporting Information). Most triple-target cliffs are formed by three large purple nodes and one large red node (code 2221) that also participate in the formation of dual-target cliffs. These compounds form prominent activity cliffs and, together with surrounding green nodes (code 0002 and 1002), make large contributions to SAR discontinuity. Furthermore, in the MT data set, a cluster is identified (cluster 3 in Figure 5b) where all compounds make large contributions to SAR discontinuity. Here, a compound represented by a blue node (code 221) forms three dual-target cliffs with different weakly potent compounds (code 000).

Such compounds that make large contributions to multitarget SAR discontinuity and also form prominent activity cliffs are a key component of multitarget activity landscapes and prime candidates for the exploration of multitarget SAR determinants. The analysis of structural features that distinguish these activity cliff markers and their activity profiles from each other, as discussed below, is of high relevance for practical applications.

Interpretation of Exemplary Activity Cliffs. Compounds forming exemplary single- and dual-target activity cliffs in the MT data set are shown in Figure 6, and compounds forming prominent activity cliffs in the other compound sets are reported in Figure S4 of the Supporting Information. In Figure 6a, exemplary MT single- and dual-target activity cliffs are displayed. The compound pair on the left forms a single-target cliff. Comparing the structures, it is evident that the removal of a carboxyl group renders a compound highly potent against all three transporters and hence nonselective. By contrast, the analogue containing the carboxyl group is selective for SHT over DA and, to a lesser extent, NE. The compound pair on the right forms a dual-target cliff. Here, the replacement of an N-substituted piperidine ring with a chemically more complex seven-membered heteroaliphatic ring system leads to a significant change in the activity profile

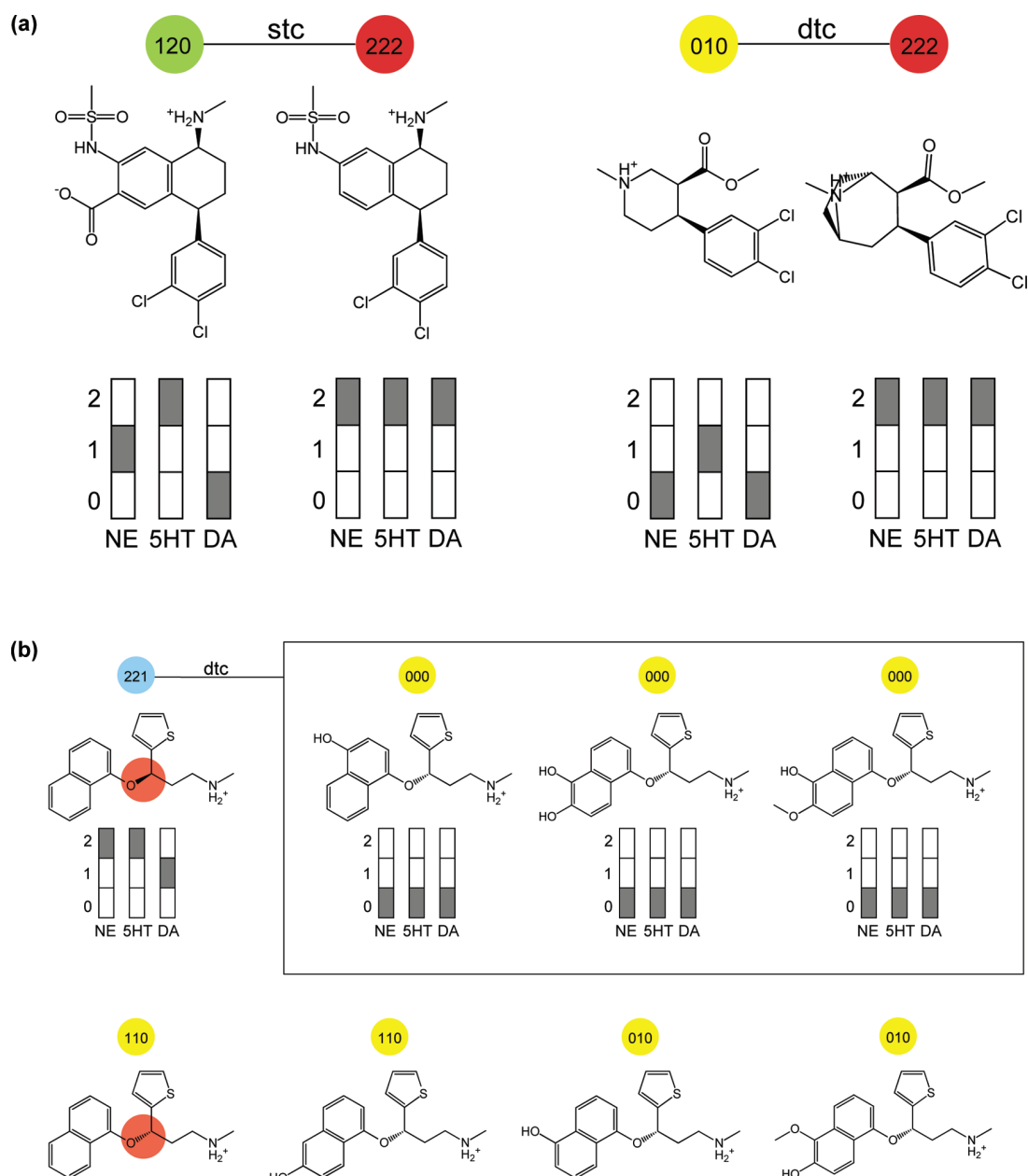


Figure 6. Exemplary activity cliffs. (a) Structures, node combinations, and activity profiles are shown for compounds forming representative single- and dual-target activity cliffs in the MT compound set. Target and activity cliff degree abbreviations are according to Table 1 and Table 2, respectively. (b) A compound involved in the formation of three dual target activity cliffs and its partner compounds are shown (see cluster 4 in Figures 4 and 5b). In addition, structures of compounds in the vicinity of these cliffs are also shown. The stereocenters of the two enantiomers with codes “221” and “110” are highlighted.

and is responsible for high potency against all three transporters. In Figure 6b, the key compound from cluster 4 in Figure 5b is shown (blue node, code 221) that is involved in the formation of three dual-target activity cliffs together with its cliff partners. Also shown are the four remaining compounds from this cluster that are moderately potent against NE and 5HT or only 5HT and not involved in the formation of activity cliffs. This compound series contains a conserved thiophene ring and a differently substituted naphthalene moiety. Comparing the structures of analogs involved in the formation of dual-target cliffs, it is evident that the weakly potent compounds differ from the highly potent one by two features, including single or dual hydroxyl (or hydroxyl and

ether) substituents at the naphthalene ring and, in addition, different chirality of a carbon atom of the “ether bridge” connecting the naphthalene and thiophene moieties. Thus, by only comparing these four analogs, it cannot be concluded with certainty which structural changes might be responsible for the dramatic reduction in potency against all three transporters, leading to the formation of dual-target cliffs. However, by inspecting the structures of compounds in the vicinity of these activity cliffs it becomes clear that the chiral center plays a major role. All consistently weakly, or weakly and moderately, potent compounds from this cluster display the same chirality, and one of these compounds is the exact enantiomer of the potent key

compound represented by the blue node. Comparing the activity profiles of these enantiomers and other compounds in the cluster, further conclusions can be drawn. The code of the highly potent enantiomer is “221”, and the code of the weakly potent one is “110”. Thus, it appears that the stereochemical switch leads to a general reduction in potency against all three transporters, whereas hydrophilic substitutions at the naphthalene ring further reduce potency in a transporter-selective manner. Thus, the analysis of this activity cliff region alone in the MT data set provides helpful information about the multitarget structure–activity relationships of these compounds. We would conclude that hydrophilic substitutions of the high-potency “221” enantiomer might be a promising route to generate analogs with differentiated MT activity profiles. The identification of key compounds that provide interpretable structure–activity profile relationship information represents a prime application for multitarget activity landscape models, as introduced herein.

Concluding Remarks. Activity landscape models of compound data sets are obtained on the basis of systematic comparisons of structural and potency relationships. Landscape models are useful computational tools for the study of SAR features contained in compound sets and the identification of key compounds that determine local or global SAR characteristics.^{1,2} For compounds active against an individual target, the generation and analysis of activity landscapes is straightforward, and different types of 2-D and 3-D representations of varying complexity have been introduced.² For compounds active against two targets, selectivity landscapes have also been generated. However, the design of multitarget activity landscapes is difficult to conceptualize and has remained an unsolved problem as of yet. Herein, we have introduced a first approach to construct multitarget activity landscapes that is based on a numerical encoding scheme of compound activity profiles derived from potency values against multiple targets. On the basis of systematic activity profile comparisons, we have derived a generally applicable formal organization of single-target and multitarget activity cliffs. The multitarget landscapes are displayed using a modified and extended version of network-like similarity graphs. In these representations, single-target and multitarget activity cliffs are easily identified. They are also formally defined by ternary code signatures that can be utilized for systematic mining of the data structure. Our multitarget activity landscape approach has been applied to characterize four compound data sets directed against three or four members of different target families. Compounds with confirmed activity annotations against more than four targets were difficult to find in public domain sources. As one might expect, compounds active against closely related protein family members often had similar activity profiles, and consequently, multitarget activity cliffs were generally more sparsely distributed than single-target cliffs. However, dual- or triple-target activity cliffs were readily identified in the activity landscapes of different data sets. In a number of instances, multitarget activity cliffs were centered on small sets of compounds that formed complex cliff patterns in multitarget graphs. The multitarget activity landscape approach and activity cliff hierarchy introduced herein is thought to provide a basis for the analysis of complex activity landscapes.

■ ASSOCIATED CONTENT

Supporting Information. Figure S1 shows multitarget activity landscapes for the AR, OR, and CA data sets. Figure S2

shows selected clusters that form activity cliffs in the AR, OR, and CA compound sets. Figure S3 shows graph representations of single-target activity cliff distributions for all four data sets. Figure S4 shows exemplary activity cliffs for AR, OR, and CA. Results contain descriptions of multitarget activity landscape representations and activity cliff patterns for these three compound sets. This information is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +49-228-2699-306; fax: +49-228-2699-341; e-mail: bajorath@bit.uni-bonn.de.

■ ACKNOWLEDGMENT

M.W. is supported by Boehringer-Ingelheim Pharma, Biberach, Germany.

■ REFERENCES

- (1) Bajorath, J.; Peltason, L.; Wawer, M.; Guha, R.; Lajiness, M. S.; Van Drie, J. H. Navigating structure–activity landscapes. *Drug Discovery Today* **2009**, *14*, 698–705.
- (2) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity landscape representations for structure–activity relationship analysis. *J. Med. Chem.* **2010**, *53*, 8209–8223.
- (3) Shanmugasundaram, V.; Maggiora, G. M. Characterizing Property and Activity Landscapes Using an Information–Theoretic Approach. Proceedings of the 222nd American Chemical Society National Meeting, Division of Chemical Information, Chicago, IL, August 26–30, 2001; American Chemical Society: Washington, DC, 2001; Abstract No. 77.
- (4) Guha, R.; Van Drie, J. H. Assessing how well a modeling protocol captures a structure–activity landscape. *J. Chem. Inf. Model.* **2008**, *48*, 1716–1728.
- (5) 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. *J. Med. Chem.* **2008**, *51*, 6075–6084.
- (6) Peltason, L.; Iyer, P.; Bajorath, J. Rationalizing three-dimensional activity landscapes and the influence of molecular representations on landscape topology and the formation of activity cliffs. *J. Chem. Inf. Model.* **2010**, *50*, 1021–1033.
- (7) Maggiora, G. M.; Shanmugasundaram, V.; Lajiness, M. S.; Doman, T. N.; Schulz, M. W. A Practical Strategy for Directed Compound Acquisition. In *Cheminformatics in Drug Discovery*; Oprea, T. L., Ed.; Wiley-VCH: Weinheim, Germany, 2005; pp 317–332.
- (8) 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.
- (9) Maggiora, G. M. On outliers and activity cliffs: Why QSAR often disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535–1535.
- (10) Peltason, L.; Hu, Y.; Bajorath, J. From structure–activity to structure–selectivity relationships: Quantitative assessment, selectivity cliffs, and key compounds. *ChemMedChem* **2009**, *4*, 1864–1873.
- (11) Wassermann, A. M.; Peltason, L.; Bajorath, J. Computational analysis of multi-target structure–activity relationships to derive preference orders for chemical modifications toward target selectivity. *ChemMedChem* **2010**, *5*, 847–858.
- (12) Lounkine, E.; Wawer, M.; Wassermann, A. M.; Bajorath, J. SARANEA: A freely available program to mine structure–activity and structure–selectivity relationship information in compound data sets. *J. Chem. Inf. Model.* **2010**, *50*, 68–78.

- (13) Fruchterman, T. M. J.; Reingold, E. M. Graph drawing by force-directed placement. *Software: Pract. Exper.* **1991**, *21*, 1129–1164.
- (14) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
- (15) ChEMBL. <http://www.ebi.ac.uk/chembl> (accessed July 1, 2010).

Summary

A first multi-target activity landscape was developed by modifying and further extending the NSG data structure to capture compounds active against multiple targets. Based on a numerical encoding scheme and systematic activity profile comparisons, a hierarchical organization of multi-target activity cliffs was derived. The approach was applied to characterize compounds active against three or four closely related targets belonging to the same target family. Accordingly, the majority of the activity cliffs were single-target cliffs. However, dual- and triple-target cliffs were also detected. My contribution to this work has been the design and analysis of the activity landscape model.

The landscape model introduced herein is in principle not limited in the number of targets. However, high-dimensional activity spaces (e.g., defined by compound activity annotations against more than 50 targets) can be difficult to navigate and comprehend using network representations. In the next study, we have introduced a first high-dimensional activity landscape that greatly reduces the complexity of bioactivity spaces and accounts for complex ligand-target relationships.

Chapter 3

Navigating High-Dimensional Activity Landscapes: Design and Application of the Ligand-Target Differentiation Map

Introduction

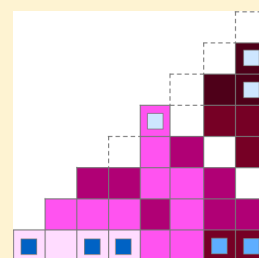
A major limitation for the design and development of activity landscapes for high-dimensional activity spaces has been the lack of publicly available profiling data. Such bioactivity spaces represent indispensable sources of activity data for pharmaceutical research, however, they are difficult to navigate and rationalize. In the following study, a first high-dimensional activity landscape, the ligand-target differentiation (LTD) map, will be introduced. Using a publicly available subset of 1496 kinase inhibitors with activity data for 172 kinases¹ the utility of the LTD map to capture complex ligand-target relationships and deconvolute compound activity patterns will be demonstrated.

Navigating High-Dimensional Activity Landscapes: Design and Application of the Ligand-Target Differentiation Map

Preeti Iyer,[†] Dilyana Dimova,[†] Martin Vogt, and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113 Bonn, Germany

ABSTRACT: The transformation of high-dimensional bioactivity spaces into activity landscape representations is as of yet an unsolved problem in computational medicinal chemistry. High-dimensional activity spaces result from the experimental evaluation of compound sets on large numbers of targets. We introduce a first concept to represent and navigate high-dimensional activity landscapes that is based on a data structure termed ligand-target differentiation (LTD) map. This approach is designed to reduce the complexity of high-dimensional bioactivity spaces and enable the identification and further analysis of compound subsets with interesting activity and structural relationships. Its utility has been demonstrated using a set of more than 1400 inhibitors with exact activity measurements for varying numbers of 172 kinases.



1. INTRODUCTION

The experimental evaluation of compounds on arrays of biological targets, often referred to as compound profiling, has become an important source of activity data for pharmaceutical research and chemical biology.¹ Compound profiling is often carried out for major therapeutic target families such as G protein coupled receptors² or protein kinases.³ In profiling campaigns, structurally diverse or, alternatively, focused compound collections are screened against varying numbers of targets. This is often (but not always) done for targets providing a representative subset of a given family. The resulting profiling data constitute high-dimensional bioactivity spaces, which are generally difficult to represent and navigate.⁴ However, in such activity spaces, ligand-binding profiles of targets, compound activity patterns, and ligand-target relationships can be explored. Furthermore, it might be attempted to identify chemical probes that differentiate between related targets⁴ or prioritize compounds for further chemical exploration and discovery efforts.³

On the basis of the activity landscape concept,^{5,6} representations of activity spaces are often generated by integrating structure and activity relationships between sets of compounds.⁶ Activity landscapes provide an intuitive access to structure–activity relationship (SAR) information but are usually focused on a specific biological activity, i.e., a single target. However, the activity landscape concept has also been extended to pairs of targets⁷ or more than two targets⁸ in order to explore the target selectivity of active compounds or the formation of multi-target activity cliffs.⁹ Only recently, multi-target activity landscape representations have been introduced, including annotated molecular networks, the original multi-target landscape design,⁸ structural similarity and activity similarity difference maps that utilize plots of compound activity versus structural similarity,¹⁰ and a landscape layout based on self-organizing maps to group structurally similar compounds together and encode their activity relationships.¹¹ However, in these representations, activities against only a few

targets (e.g., three or four) can be captured in a meaningful and interpretable way. The representation of high-dimensional activity spaces (e.g., involving 50, 100, or more targets) in an activity landscape format has thus far not been reported.

The design of high-dimensional activity landscapes has been hampered by the limited availability of compound profiling data in the public domain. Although a number of pharmaceutical companies have already generated large bodies of profiling data for popular therapeutic targets, most of this data is kept proprietary, for understandable reasons. A notable exception has been a recent study by a group from Abbott Laboratories.¹² For the generation of kinase interaction networks and the exploration of polypharmacology patterns, a total of 3858 compounds were tested against varying numbers of 172 kinases representing a diverse sample of the kinome.¹² As a part of this investigation, structures and activity data for a subset of 1496 of these compounds were made publicly available, hence providing a significant source of profiling data for further studies. Using this data set, we have developed and applied a first concept for the design and analysis of high-dimensional activity landscapes that is reported herein.

2. ACTIVITY DATA

From the publicly released Abbott data set, all compounds with unique 2D molecular representations¹³ were extracted for which a K_i value for at least one kinase was available, leading to the selection of 1473 compounds. These compounds were annotated with pK_i values for one to 122 kinases. The activity annotations included all 172 kinases investigated in the Abbott study. A maximum of 101 kinases were shared between individual compounds. In our analysis, only absolute equilibrium constants were considered as activity measurements and threshold measurements were ignored. For activity landscape

Received: May 9, 2012

Published: July 14, 2012

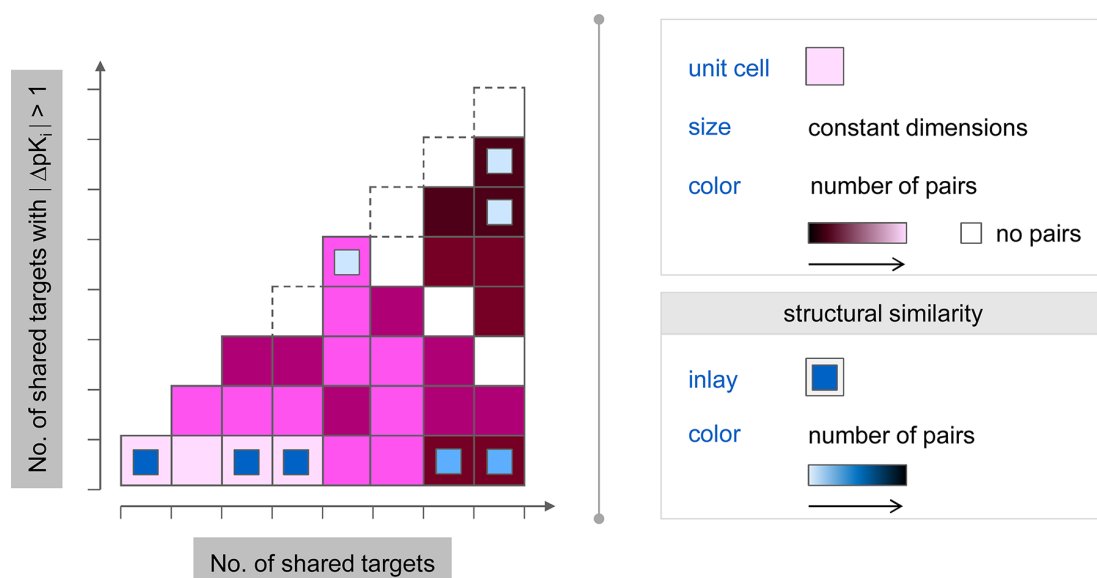


Figure 1. Design principles of the ligand-target differentiation map. The schematic illustration summarizes the basic design elements of the ligand-target differentiation map. Unit cells have constant dimensions and delineate a well-defined range of shared targets (x -axis) and of shared targets with qualifying potency differences (y -axis). Compound pairs are assigned to a unit cell if the underlying target relationship defined by the cell is met. Color coding accounts for the number of qualifying compound pairs from light pink (many pairs) over magenta to black (a single pair). Inlay squares indicate structural relationships and are “inversely” color-coded according to the frequency of structurally related pairs in a unit cell from black (many pairs) over dark blue to light blue (a single pair).

design, this data set presented a challenging test case because the underlying profiling matrix was high-dimensional yet incomplete, i.e., compounds were assayed against varying numbers of targets and activity profiles only partly overlapped in many instances. The activity profile of a compound consists of all of its target annotations.

3. LIGAND-TARGET DIFFERENTIATION MAP

On the basis of our evaluation, annotated molecular network representations, which were previously utilized for the generation of single- and multi-target activity landscapes, were not suitable for capturing and representing high-dimensional activity spaces. The design of high-dimensional activity landscapes presents challenges that go beyond the analysis of single-target SARs⁶ or multi-target SAR discontinuity patterns.^{6,8} In particular, ligand-target relationships need to be systematically explored and compared in light of structural features of active compounds. Therefore, it was required to investigate new representation concepts. In the following, the basic design principles of the Ligand-Target Differentiation (LTD) map, our central data structure for high-dimensional activity landscape analysis, and its elements are discussed. In addition, the extraction of compound and activity information from the map is illustrated.

3.1. Design Concept. As an activity landscape representation, the LTD map must systematically account for compound potency and similarity relationships in high-dimensional data sets. A major goal of such representations is to provide complete coverage of experimental data. For the analysis of compound profiling data, it must be considered that high-dimensional matrices are often incomplete. Hence, the graphical data structure must be flexible and capable of capturing profiling matrices of different composition.

Figure 1 shows a schematic representation of an LTD map to illustrate its design principles. The LTD map is based on four general principles:

- (1) The basic unit of the data structure is a compound pair.
- (2) Compounds are differentiated according to the number of targets they share.
- (3) Compounds are further differentiated according to their activity differences against these targets.
- (4) Structural relationships between all molecules are monitored.

Accordingly, all pairwise target, activity profile, and structural relationships between active compounds are initially determined. The LTD map then relates the number of targets shared by any pair of compounds to the number of targets against which these two compounds display significant differences in potency using a “unit cell” as its basic data element. In our analysis, the threshold for the potency difference between compounds in a pair against a common target is set to 1 order of magnitude (a flexible criterion, depending on data set characteristics). The relative frequencies of detected target and structural relationships are then captured through color coding and map annotation. Figure 1 summarizes the basic elements of the LTD map, and Figure 2 shows the LTD map representation of the entire kinase inhibitor data set, as further discussed in the following.

3.2. Elements of Graphical Representation. In the map, a unit cell is represented as a square. The constant dimensions of a unit cell in Figure 2 are five by five. The numbers of shared targets and shared targets with more than an order of magnitude potency difference are reported along the x -axis and the y -axis, respectively. The LTD map thus consists of an array of squares that account for all pairwise target and potency difference relationships in a data set and span the entire ranges. The number of compound pairs falling into each cell is

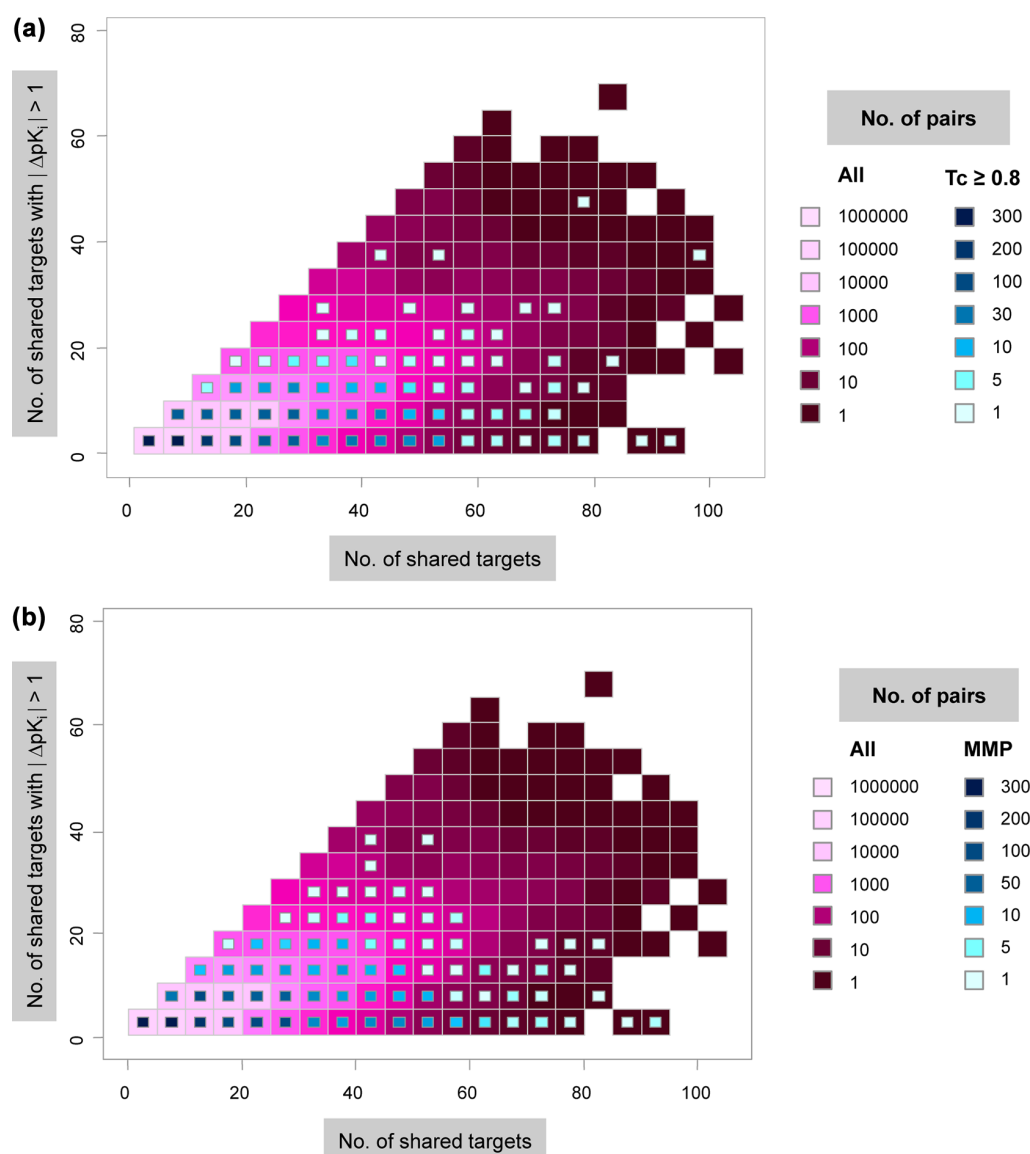


Figure 2. LTD map of kinase inhibitor data. Two versions of the LTD map of the kinase inhibitor data set are shown that only differ in the way structural relationships between active compounds are determined. In panel (a), MACCS Tanimoto similarity relationships are shown, and in panel (b) MMP-based substructure equivalences are shown (i.e., cells with inlays contain pairs of compounds with a common core structure).

determined, and a continuous color code (from black over magenta to light pink) is used to monitor the frequency of relationships within the cells (i.e., a cell colored in black contains a single pair, and a cell in light(est) pink contains the maximally observed number of pairs). An “empty” cell within the map indicates that no compound pairs are falling into the respective data intervals.

Structural relationship information is also incorporated into the LTD map. For our analysis, compound similarity was assessed in two complementary ways. Pairwise whole molecule Tanimoto similarity¹⁴ was calculated using MACCS structural keys.¹⁵ As a similarity threshold for selected compound relationships, a Tanimoto coefficient of 0.8 was applied. Furthermore, matched molecular pair (MMP) analysis^{16,17} was carried out to identify substructure relationships between compounds. For this purpose, all compound pairs were identified that formed an MMP, i.e., that shared a given key

fragment (core structure), as described previously.¹⁷ It should be noted that increasing the potency and similarity thresholds decreases the number of qualifying compound pairs and hence the information content of the analysis. Data noise and information content must be balanced.

All unit cells that contain compound pairs with structural relationships are then marked through the addition of “square inlays”, as schematically illustrated in Figure 1. Furthermore, Figure 2a captures structural relationships between the kinase inhibitors on the basis of pairwise Tanimoto similarity calculations and Figure 2b on the basis of common (MMP-based) core structures. An inversely shaded color code (from light blue over dark blue to black) is applied to the inlays in order to account for the frequency of pairwise structural relationships per cell (i.e., a cell with a light(est) blue square inlay contains a single relationship). Hence, annotation of cells

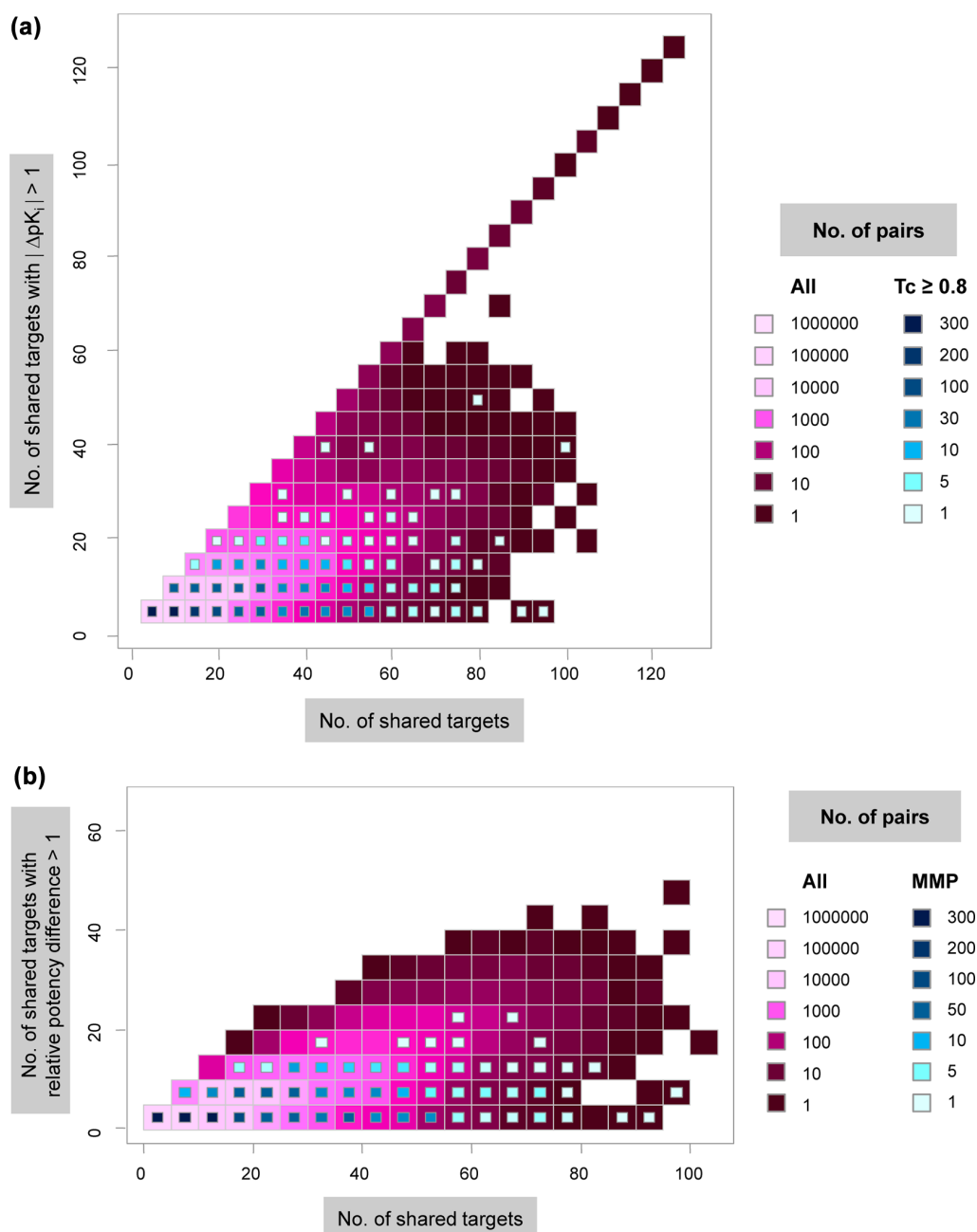


Figure 3. Data set and map modifications. In panel (a), the LTD map of the kinase inhibitor data set was calculated after addition of a hypothetical compound with 1 mM potency against all 172 kinases. In panel (b), an alternative version of the LTD map is shown for the original data set where potency deviations from mean compound potency were used instead of absolute potency differences, as rationalized in the text.

monitors the distribution of structural relationships in high-dimensional activity space.

A consistent numbering scheme is applied to cells in LTD maps, adhering to “top down” (vertical) followed by “from left to right” (horizontal) reading directions. Thus, if multiple cells have the same number of targets with significant potency differences, they are numbered in the order of increasing numbers of shared targets. LTD maps were drawn using routines implemented in the R environment.¹⁸

3.3. Interpretation. The LTD map provides an immediate view of the data distribution in high-dimensional activity space, as illustrated in Figures 1 and 2. For the kinase inhibitor data

set containing activities against a total of 172 different kinases, individual compound pairs share up to 101 kinases and up to 69 kinases with potency differences of more than 1 order of magnitude. As clearly delineated by the cell color code, the bulk of the activity data falls into the map section spanned by zero to ~20 shared targets and zero to ~10 targets with significant potency differences. In addition, the section of highly populated cells extends to ~60 shared targets and ~30 targets with potency differences. For further increasing numbers of shared targets and targets with qualifying potency differences, the number of compound pairs rapidly declines. The inlay view of structural relationships between compound pairs adds further

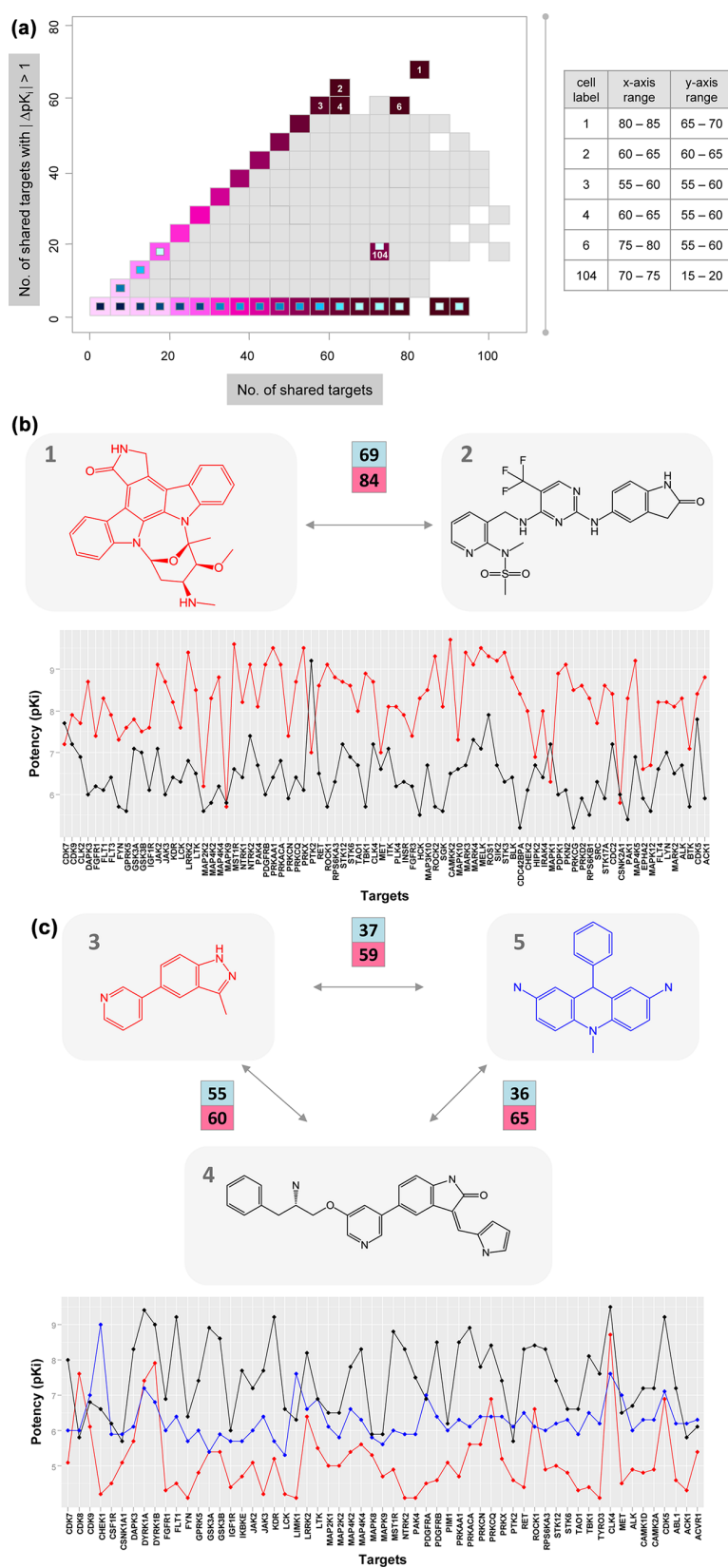


Figure 4. continued

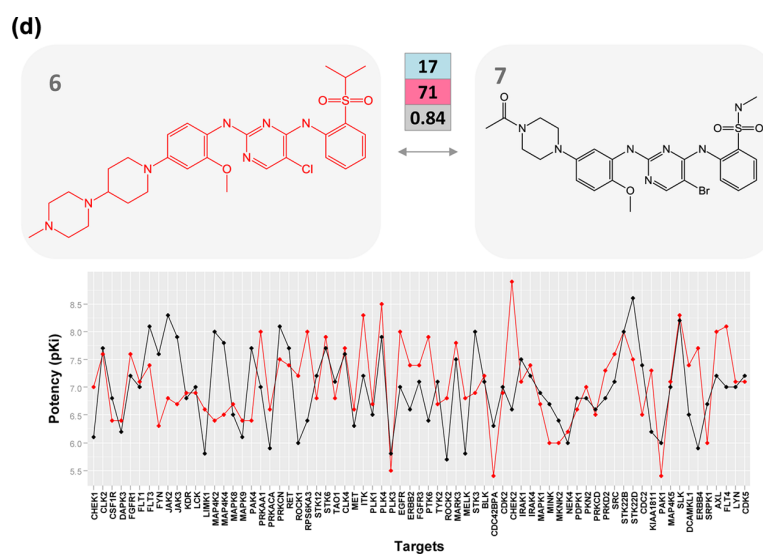


Figure 4. Compound information. The figure illustrates how compound information is extracted from the LTD map. In panel (a), the LTD map in Figure 1a is displayed in simplified form. Map boundaries referred to in the text and exemplary cells are color-coded while other regions are displayed in light gray. In addition, selected cells are numbered (following the numbering scheme described in the text). The table insert reports the x -axis and y -axis range for each labeled cell. In panels (b)–(d), compounds taken from selected cells are shown together with their activity profiles. Cells from which the compounds originate are specified. For pairwise comparisons, the total number of shared kinases and the subset of kinases with significant potency differences are reported on a pink and light blue background, respectively. In addition, for compound pairs with MACCS Tanimoto similarity above the threshold value, the Tanimoto coefficient is given on a gray background. In the activity profiles, pK_i values for all shared targets are reported. Compounds and corresponding profiles are color-coded. Kinase abbreviations are used according to ref 12.

information to the map. Cells are highlighted that contain compound pairs with structural relationships. In Figure 2a and b, Tanimoto similarity relationships and (MMP-based) substructure equivalences are displayed, respectively. These figures show how structural similarity relationships are distributed in high-dimensional activity space. In the kinase inhibitor data set, the distributions of Tanimoto similarity and substructure relationships are similar, as revealed in Figure 2. Most similarity relationships are detected between compounds that share only a few targets. Frequent similarity relationships still occur between compounds sharing up to ~ 50 targets and ~ 10 targets with significant potency differences (dark to medium blue inlay area in Figure 2). For increasing numbers of shared targets and targets with potency differences, only few structural relationships between inhibitors are detected. The bottom row of cells in the map contains compound pairs that have similar activity against all kinases against which they have been tested. This region contains many structurally similar compounds (as one might expect). By contrast, compound pairs forming the leftmost (pseudo-diagonal) cell layer display potency differences against all, or nearly all, of their shared targets.

On the basis of the information provided by the LTD map, activity data can be further analyzed by selecting compounds from map segments or individual cells of interest. In the following, representative examples are discussed.

3.4. Modifications of LTD Maps. Compounds forming pairs on the diagonal of the LTD maps include those that display differential activity against kinases. In addition, cells on the diagonal might also include compounds having consistently higher or lower potency against shared targets. These compounds are less interesting for further analysis. As an extreme case, a data set might contain one or more compounds with artificially high or low potency against many targets, which

would result in the formation of many artificial compound pairs and complicate the analysis of the LTD map. Such compounds were not present in the kinase inhibitor set. However, if such compounds exist in a data set, they will exclusively form cells on the diagonal, which alters the map appearance in a characteristic manner. This is demonstrated in Figure 3a that shows the LTD map of the kinase inhibitor data set recalculated after addition of a hypothetical compound with 1 mM potency (pK_i value of 3) against all 172 kinases. Because such compounds only occur in cells on the diagonal, they (and the pairs they form) can be easily identified and removed from further analysis.

However, to principally omit this potential complication, the LTD map can be modified by considering potency differences with respect to the average potency of compounds for the set of shared targets instead of absolute potency differences. Thus, for each compound in a pair, the average potency for its shared targets is calculated and subtracted from each individual potency value, which yields relative potency values. These values reflect whether a compound shows above or below average potency values for its shared targets. The differences between these relative potency values are then used for map construction. The modified LTD map calculated for the potency difference threshold of 1 order of magnitude as before is shown in Figure 3b. The diagonal cells are less densely populated, which is due to the fact that the median of relative potency differences is only 0.54 compared to 0.80 for absolute differences. Compounds with artificially high or low potencies against many targets no longer form pairs with large relative potency differences and do not populate diagonal cells in this map. Only compounds with differentiated potency profiles (i.e., compounds with selectivity) can induce signals and form pairs that populate prominent cells. Accordingly, the modified LTD map in Figure 3b remained essentially constant when it was recalculated after addition of the hypothetical compound.

3.5. Alternative Potency Difference Threshold Values.

It should also be mentioned that potency difference threshold values can not only be prespecified but can also be determined in a meaningful manner on the basis of the potency distributions within a given data set, as demonstrated in the following for the kinase inhibitor set. In order to evaluate the significance of potency differences for a target observed for pairs of compounds, the distribution of potency differences for all pairs of compounds and all shared targets was analyzed. Thus, for all 786,776 compounds pairs, the potency differences for one to 101 targets shared between them were pooled. Observed potency differences ranged from zero to a maximum of 7.1. The mean value was 1.0, with a standard deviation of 0.8. These values reflected the asymmetric nature of the distribution, with values extending from about one standard deviation below the average to more than seven standard deviations above the average. This was also indicated by the median of 0.8 and the interquartile range with a first quartile of 0.4 and a third quartile of 1.5. On the basis of these values, a potency difference threshold of 1 order of magnitude applied in our analysis was a reasonable choice for this data set because it directed the analysis toward compound pairs with a high number of above average potency differences.

On the basis of these considerations, a statistical analysis of the significance of compound pairs with a certain number of targets with above average potency difference with respect to the total number of commonly annotated targets might be performed. For example, by taking the median potency difference as a threshold value, half the potency differences between a pair of compounds would be expected to have values beyond the threshold. This gives rise to a binomial distribution with $p = 0.5$. In this case, significance at the 0.01 level would correspond, for instance, to compound pairs with 15 large potency differences given a total of 20 shared targets. These values can guide the analysis of the LTD map in order to identify interesting compound pairs especially considering the incompleteness of data set annotations.

4. COMPOUND DATA ANALYSIS

Figure 4a points at regions in the LTD map in Figure 2 that contain interesting compounds for further exploration. Pairs can be automatically extracted from cells. The consistent numbering scheme of cells in LTD maps is also illustrated in Figure 4a. Cell 1 contains a pair of structurally distinct inhibitors that share 84 targets and yield significant potency differences for 69 of them. This represents the largest number of shared targets with qualifying potency differences detected within the data set. The structures of these compounds and their activity profiles are shown in Figure 4b. Compound 1 has mostly higher potency than compound 2, which explains the overall large number of potency differences (a situation observed in a number of instances in this region of the map). However, both compounds display different activity profiles and significant differentiation potential against many kinases, with in part large differences in potency, especially in the case of compound 1. Furthermore, the adjacent cells 2–4 contain compound pairs active against ~60 kinases with significant potency differences against many of them. Figure 4c shows three exemplary compounds taken from these cells. Compounds 3 and 4 form a pair and have different potencies against 55 of the 60 targets they share. As an additional example, compound 5 is included in the comparison. Compound 3 has overall lower potency than compound 4 but the traces of their

activity profiles show notable similarities. Compound 4 has high kinase differentiation potential, often with relative differences in potency between kinases of 3 orders of magnitude or more. By contrast, compound 5 has mostly intermediate potency and shows rather limited ability to differentiate between kinases. Cell 104 in Figure 4a maps to another interesting region in the LTD map. In this region, compounds have similar activity against many kinases and yield only a limited number of significant potency differences. Figure 4d shows a pair of structurally similar compounds taken from cell 104. Their activity profiles are overall also similar but reveal a number of notable potency differences against individual kinases. Thus, the comparison of compounds with many shared targets that include only a limited number of targets with significant potency differences might identify potential selectivity probes. Taken together, these examples illustrate how compound information can be extracted from the LTD map and how compound subsets with desired properties can be selected for further studies.

5. CONCLUDING REMARKS

We have introduced the ligand-target differentiation map that is designed to navigate high-dimensional bioactivity spaces taking structural relationships between active compounds into account. As such, the LTD map represents a high-dimensional activity landscape. A hallmark of the activity landscape concept is the graphical integration of compound structure and activity relationships. One of the difficulties involved in exploring the design of high-dimensional landscapes has been the limited availability of relevant compound profiling data, at least in the public domain. A notable exception is provided by the publicly available kinase inhibitor data set from Abbott that contains activity data for more than 1400 inhibitors and 172 different kinases and probably represents the largest high-dimensional compound profiling set currently available in the public domain. Our exploratory efforts have been much supported by the availability of this data set, leading to the design of the LTD map structure. A key feature of the newly introduced approach is the reduction of the inherent complexity of variable high-dimensional activity spaces. This is accomplished by systematically accounting for pairwise differences between multi-target activity profiles of test compounds. Such differences are graphically represented by monitoring the number of common targets with significant potency differences as a function of the total number of targets that are shared by compound pairs. This representation greatly simplifies the navigation of high-dimensional activity spaces and makes it possible to quickly focus on regions in data sets that are most interesting for further analysis. Another key feature of the LTD map is its basic data element, a constantly sized cell that contains compound pairs with well-defined relationships. Through color coding and cell annotation, both activity and structural relationship information is provided. From individual cells or groups of cells, compound pairs and subsets with well-defined relationships can be selected, as demonstrated herein. The generation of the LTD map is straightforward, and the representation is also applicable to smaller data sets of lower dimensionality. Thus, LTD maps should be helpful for many applications in multi-target compound data analysis. In addition, it is hoped that the approach introduced herein might also catalyze the development of alternative high-dimensional activity landscape views.

AUTHOR INFORMATION

Corresponding Author

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

Author Contributions

†The contributions of these two authors should be considered equal.

Notes

The authors declare no competing financial interest.

REFERENCES

- (1) Rix, U.; Superti-Furga, G. Target profiling of small molecules by chemical proteomics. *Nature Chem. Biol.* **2009**, *5*, 616–624.
- (2) Allen, J. A.; Roth, B. L. Strategies to discover unexpected targets for drugs active at G protein-coupled receptors. *Annu. Rev. Pharmacol. Toxicol.* **2011**, *51*, 117–144.
- (3) Goldstein, D. M.; Gray, N. S.; Zarrinkar, P. P. High-throughput kinase profiling as a platform for drug discovery. *Nature Rev. Drug Discov.* **2008**, *6*, 391–397.
- (4) Bajorath, J. Computational analysis of ligand relationships within target families. *Curr. Opin. Chem. Biol.* **2008**, *12*, 352–358.
- (5) Bajorath, J.; Maggiora, G.; Lajiness, M., organizers. The Emerging Concepts of Activity Landscapes and Activity Cliffs and Their Role in Drug Research; 240th National Meeting of the American Chemical Society, Divisions of Chemical Information and Computers in Chemistry, Boston, MA, August 22–26, 2010.
- (6) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity landscape representations for structure-activity relationship analysis. *J. Med. Chem.* **2010**, *53*, 8209–8223.
- (7) Peltason, L.; Hu, Y.; Bajorath, J. From structure-activity to structure-selectivity relationships: Quantitative assessment, selectivity cliffs, and key compounds. *ChemMedChem* **2009**, *4*, 1864–1873.
- (8) Dimova, D.; Wawer, M.; Wassermann, A. M.; Bajorath, J. Design of multi-target activity landscapes that capture hierarchical activity cliff distributions. *J. Chem. Inf. Model.* **2011**, *51*, 256–288.
- (9) Stumpfe, D.; Bajorath, J. Exploring activity cliffs in medicinal chemistry. *J. Med. Chem.* **2012**, *55*, 2932–2942.
- (10) Medina-Franco, J. L.; Yongye, A. B.; Perez-Villanueva, J.; Houghten, R. A.; Martinez-Mayorga, K. Multi-target structure-activity relationships characterized by activity-difference maps and consensus similarity measures. *J. Chem. Inf. Model.* **2011**, *51*, 2427–2439.
- (11) Iyer, P.; Bajorath, J. Representation of multi-target activity landscapes through target pair-based compound encoding in self-organizing maps. *Chem. Biol. Drug Des.* **2011**, *78*, 778–786.
- (12) Metz, J. T.; Johnson, E. F.; Soni, N. B.; Merta, P. J.; Kifle, L.; Hajduk, P. J. Navigating the kinome. *Nature Chem. Biol.* **2011**, *7*, 200–202.
- (13) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
- (14) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
- (15) *MACCS Structural Keys*; Symyx Software: San Ramon, CA, USA.
- (16) Kenny, P. W.; Sadowski, J. Structure modification in chemical databases. In *Cheminformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2005; pp 271–285.
- (17) Wassermann, A. M.; Bajorath, J. Large-scale exploration of bioisosteric replacements on the basis of matched molecular pairs. *Future Med. Chem.* **2011**, *3*, 425–436.
- (18) *R: A Language and Environment for Statistical Computing*; R Development Core Team, R Foundation for Statistical Computing: Vienna, Austria, 2008.

Summary

Herein, the first activity landscape to navigate high-dimensional bioactivity spaces was introduced. The newly designed LTD map had two key features. First, it reduced the complexity of high-dimensional activity spaces by accounting for pairwise potency differences between compounds. Second, it provided a compound pair-based analysis, i.e., the unit element of the landscape was a pair of compounds. Pairwise activity profile comparisons were graphically represented by monitoring the number of targets with significant potency differences as a function of the total number of shared targets between a pair of compounds. Furthermore, a color code was applied to account for pairwise structural relationships. Taken together, the LTD map greatly facilitated the analysis of multi-target compound data and enabled to quickly focus on interesting compound subsets for further exploration. My contributions to this study have been the design of the activity landscape model and its analysis.

In this and the previous study, the concept of activity landscapes has been extended to capture complex ligand-target relationships and account for multi-target SARs. Cardinal features of landscape models are activity cliffs (i.e., pairs of compounds having a significant potency difference). Thus far, activity cliff distributions have been studied on a per-target basis, and the global distribution of activity cliffs has been unknown. In the next study, a systematic survey has been carried out to analyze the cliff frequency and distribution on a large scale.

Chapter 4

Comprehensive Analysis of Single- and Multi-Target Activity Cliffs Formed by Currently Available Bioactive Compounds

Introduction

The systematic exploration of activity cliffs has experienced increasing interest in practical medicinal chemistry and drug development efforts. Activity cliffs are thought to contain high SAR information content and are cardinal features of activity landscape models. On the basis of landscape representations, single- and multi-target cliffs have been detected in various activity classes. Thus far, the distribution of activity cliffs across different biological targets has not been reported. Therefore, we have systematically searched for single- and multi-target activity cliffs in major public domain repositories to statistically account for their global distribution.

Comprehensive Analysis of Single- and Multi-Target Activity Cliffs Formed by Currently Available Bioactive Compounds

Anne M. Wassermann[†], Dilyana Dimova[†]
and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität Bonn, Dahlmannstr. 2, D-53113 Bonn, Germany
*Corresponding author: Jürgen Bajorath, bajorath@bit.uni-bonn.de
[†]The contributions of these authors should be considered equal.

Activity cliffs are formed by structurally similar compounds having large potency differences. Their study is a focal point of SAR analysis. We present a first systematic survey of single- and multitarget activity cliffs contained in currently available bioactive compounds. Approximately 12% of all active compounds were involved in the formation of activity cliffs. Perhaps unexpectedly, activity cliffs were found to be similarly distributed over different protein target families. Moreover, only approximately 5% of all activity cliffs were multitarget cliffs. Importantly, we also found that only very few multitarget cliffs were formed by compounds having different target selectivity. In addition, 'polypharmacological cliffs', i.e., multitarget activity cliffs involving targets from different protein families, were also only rarely found. Taken together, our findings reveal that only approximately 2% of all pairs of structurally similar compounds sharing the same biological activity form activity cliffs but that, on average, approximately one of 10 active compounds is involved in the formation of one or two single-target cliffs of large magnitude (with at least 100-fold difference in potency). These compounds provide a rich source of SAR information and can be identified across many different target families.

Key words: activity cliffs, activity landscapes, compound potency, data mining, polypharmacology, structure–activity relationships, target families

Received 31 March 2011, revised 18 May 2011 and accepted for publication 19 May 2011

In the context of SAR analysis, the concepts of activity landscapes (1,2) and activity cliffs (2,3) have experienced increasing interest in recent years (4). Activity landscapes can be rationalized as graphical representations that integrate molecular similarity and potency

relationships between active compounds (2), and activity cliffs represent their most prominent features (2,3). An activity cliff is formed by structurally similar compounds having large differences in potency (2,3). Therefore, activity cliffs represent the extreme form of SAR discontinuity (1,2) and their presence in compound sets is often responsible for difficulties in deriving QSAR models for activity prediction (3). Given their 'small structural change – large potency effect' phenotype, activity cliffs are also regarded as the most informative feature of landscape models (1–3), both from a medicinal chemistry (1,2) and from a chemoinformatics (5,6) perspective. For chemical optimization efforts, the assessment of activity cliffs in compound series plays an important role.

Activity cliffs have conventionally been analyzed for compounds active against individual targets ('single-target cliffs'). However, for compounds with activity against multiple targets, 'multitarget cliffs' might also occur (7). Such cliffs result from differential potency of a compound pair against two or more related targets. These targets might be closely related, i.e., members of the same protein family, or unrelated, if the cliff-forming compounds display polypharmacological behavior (8–10).

Although activity cliffs are intensely studied, it is currently unknown how they are globally distributed across available bioactive compounds and protein targets. Studies reported thus far have focused on identifying activity cliffs in individual compound sets, but no systematic assessment of activity cliff distributions has been carried out. Furthermore, it is currently unknown how frequently multitarget activity cliffs might actually occur in bioactive compounds. Therefore, we have systematically searched for single- and multitarget activity cliffs in public domain compounds with reported activity against human target proteins.

Materials and Methods

The three major public domain compound repositories were analyzed, i.e., PubChem,^a BindingDB (11), and ChEMBL (12). PubChem bioassays contain high-throughput screening data, whereas BindingDB and ChEMBL predominantly contain compounds taken from the medicinal chemistry literature, mostly originating from chemical optimization efforts. The latest version of BindingDB has integrated the ChEMBL compound collection for defined protein targets. Thus, we have analyzed these compounds together, in addition to PubChem compounds.

For our analysis, we extracted small compounds from public domain repositories that contained at least five heavy atoms and had a molecular weight of not more than 900 Da. A total of 164 165 unique BindingDB/ChEMBL compounds were obtained (approximately 140 000 of which originated from ChEMBL) that were reported to be active against 1355 non-redundant individual human targets. These compounds yielded 330 526 defined activity annotations (i.e., many compounds were active against multiple targets). From PubChem, 187 confirmatory inhibition assays for human targets were extracted that contained 21 532 active compounds with 30 805 defined annotations against 98 different targets. Only K_i or IC_{50} values were considered as activity annotations.

Pairwise compound similarities were calculated using the 'extended connectivity fingerprint with bond diameter 4' (ECFP4) (13) as implemented in Pipeline Pilot.^b As a similarity threshold for activity cliff formation, an ECFP4 Tanimoto coefficient (Tc) value of 0.55 was applied. This ECFP4 Tc threshold value identifies compounds with high structural similarity (14). For comparison, the calculations were also carried out with another molecular representation, the MACCS structural key fingerprint.^c The same number of pairs of similar compounds above the ECFP4 Tc threshold value of 0.55 was obtained for the MACCS Tc calculations when a threshold value of 0.85 was applied.

The targets in our analysis were grouped into target families based on the family organization of UniProt (15). Targets belonging to the G-protein-coupled receptor 1 family were divided into smaller groups following the protein classification hierarchy of ChEMBL.

Our analysis was carried out with in-house generated Perl and Java programs.

Results and Discussion

We found 164 165 BindingDB/ChEMBL compounds yielding 330 526 activity annotations against 1355 targets and 21 532 PubChem compounds yielding 30 805 activity annotations against 98 targets that met our selection criteria.

Activity profiles and cliffs

Following compound selection, we generated activity profiles (7) for all compounds using a constant representation scheme. An activity profile of a compound consisted of binned activity measurements for all annotated targets. Potency values were assigned to three different ranges, i.e., 'weakly potent' ($pK_i \leq 5$; bin label '0'), 'moderately potent' ($pK_i > 5$ and $pK_i \leq 7$; bin label '1'), or 'highly potent' ($pK_i > 7$; bin label '2'). Accordingly, a compound was considered weakly potent against a given target if its potency was lower than or equal to 10 μM , moderately potent if the potency was higher than 10 μM but lower than or equal to 100 nM, and highly potent if it was higher than 100 nM. If multiple measurements were available, a compound was only included in the analysis if all values fell into the same potency bin.

For our analysis, we applied the definition that an activity cliff was formed by a pair of compounds that exceeded the cliff similarity

threshold and in which one compound was highly potent against a given target and the other only weakly potent (i.e., representing a '2' versus '0' potency bin combination against the target). Compounds active against multiple targets can form single- or multitarget activity cliffs. The latter are termed dual-, triple-, quadruple-target cliffs, etc. according to the number of targets for which cliffs occur. In Figure 1A,B, exemplary compound pairs are shown with their activity profiles that form a single- and dual-target cliff, respectively.

Activity cliff distribution

After generating compound activity profiles, we systematically searched for single- and multitarget activity cliffs. In Table 1, we report the activity cliff distribution for BindingDB/ChEMBL compounds when, as an approximation, both K_i and IC_{50} values were considered as potency annotations. In this case, we detected a total of 36 063 single-, 1654 dual-, and 233 triple-target activity cliffs. The number of multitarget cliffs of higher degrees (target numbers) rapidly declined, although cliffs involving up to seven targets were detected. Table 1 also reports the corresponding activity cliff distribution when only directly comparable K_i values were considered as measurements. Then, 10 063 single-, 330 dual-, and 61 triple-target activity cliffs were detected (i.e., approximately one-fourth of the cliffs found when both K_i and IC_{50} values were considered). In addition, in this case, 17 cliffs involving four targets and two cliffs involving five targets were identified. For the K_i/IC_{50} - and K_i -based distributions, we found that 20 473 and 6194 of 164 165 and 56 795 compounds, respectively, were involved in the formation of activity cliffs, i.e., approximately 12.5% and approximately 10.9%.

We also determined the total number of compound pairs that could potentially form activity cliffs, i.e., pairs that exceeded the activity cliff similarity threshold. When both K_i and IC_{50} measurements were considered, 1 530 493 qualifying compound pairs yielded a total of 38 045 (single- and multitarget) cliffs. Hence, only approximately 2.5% of all qualifying compound pairs formed activity cliffs and, accordingly, approximately 97.5% did not. Furthermore, only 5.2% of all cliff-forming compound pairs represented multitarget cliffs. When only K_i values were considered, 574 851 compound pairs were found that yielded a total of 10 473 cliffs, i.e., only approximately 1.8% of these pairs formed activity cliffs and only 3.9% of these were multitarget cliffs. Thus, activity cliffs were only sparsely distributed among pairs of structurally similar compounds. Control calculations using the MACCS fingerprint yielded similar statistics; for example, for the K_i/IC_{50} -based analysis, 3.2% of all qualifying compound pairs were found to form activity cliffs and only 4.3% of these compound pairs formed multitarget cliffs. For the K_i -based analysis, the corresponding values were 2.5% and 3.8%, respectively.

In addition, for each cliff compound in a ligand set, all activity cliffs formed with its ECFP4-derived structural neighbors and the ratio of cliffs versus the number of neighbors were determined. Based on these calculations, the frequency of cliff formation in qualifying pairs of compounds involving a cliff compound was on average approximately 18% (for both the K_i - and K_i/IC_{50} -based analyses).

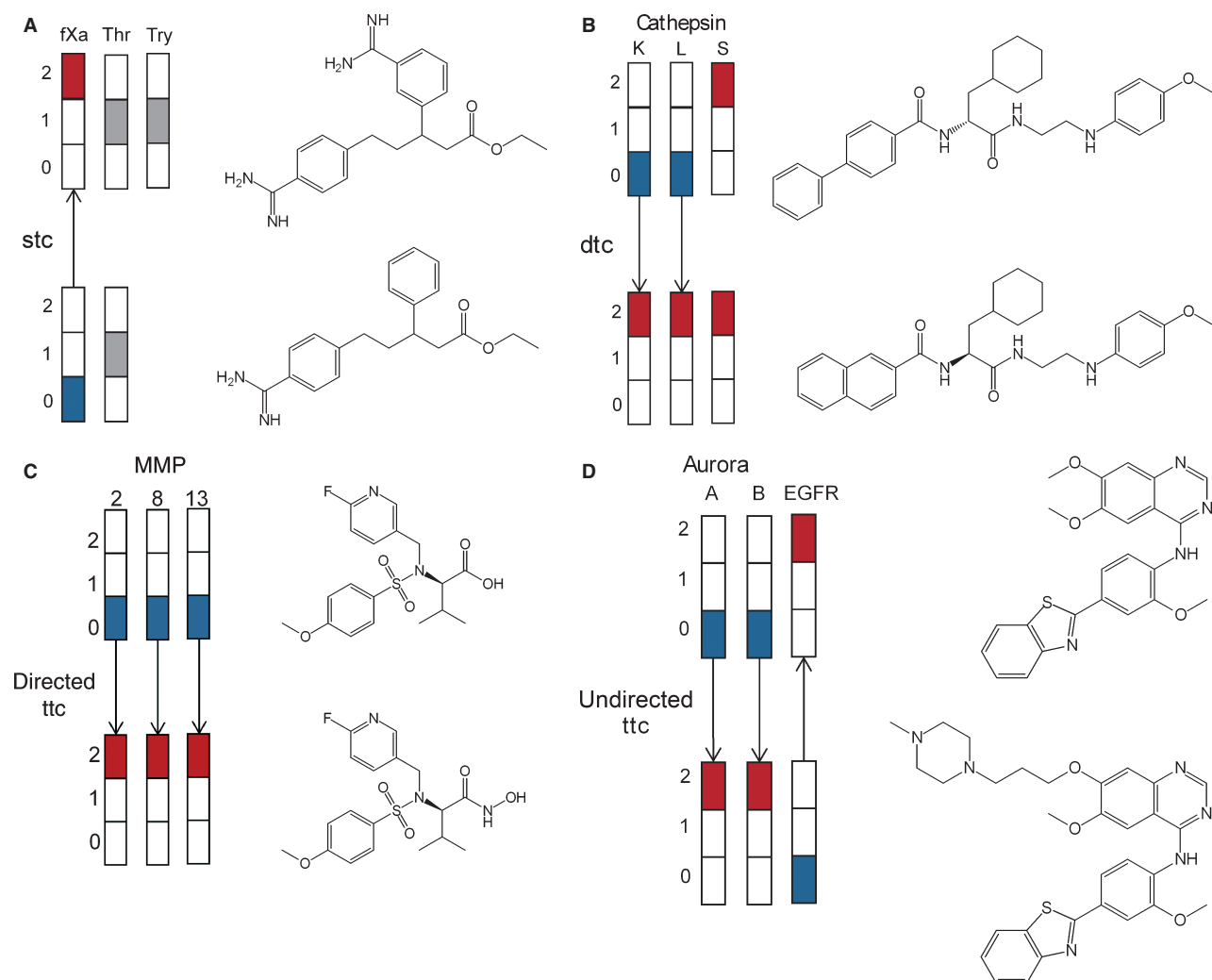


Figure 1: Exemplary single- and multi-target activity cliffs. (A) Shown are two compounds and their activity profiles. One of these compounds is active against all three, and the other against two of the serine proteases factor Xa (fXa), thrombin (Thr), and trypsin (Try). Based on their activity profiles, these two compounds form a single-target activity cliff (stc) for factor Xa. The cliff is indicated by an arrow. (B) Shown are two compounds that are active against cathepsins K, L, and S. One compound is highly potent against all three cathepsins, whereas the other is only highly potent against cathepsin S, but weakly potent against cathepsins K and L. Hence, these compounds form a dual-target cliff (dtc) for cathepsins K and L, indicated by arrows. (C) A compound pair is shown that forms a triple-target cliff (ttc) for matrix metalloproteases (MMP) 2, 8, and 13. The potency of one compound is consistently high against all three targets, and the potency of the other is consistently low, i.e., this cliff is directed. (D) A compound pair is shown that forms a triple-target cliff for Aurora serine/threonine kinases A and B and the epidermal growth factor receptor (EGFR) tyrosine kinase. In this case, the compounds show differential target selectivity. One compound is highly potent against Aurora kinases A and B and weakly potent against the EGFR kinase, whereas the other compound displays an inverse activity profile. Accordingly, this triple-target cliff is undirected.

Our active compound pool extracted from PubChem amounted to approximately one-seventh of the size of BindingDB, and these screening hits had overall lower potency than BindingDB/ChEMBL compounds, as expected. When K_i and IC_{50} values were taken into account, only 13 single-target and no multitarget cliffs were detected. These 13 activity cliffs involved only five different targets. Thus, the occurrence of activity cliffs in PubChem compounds was negligible. Hence, the further analysis of activity cliffs was limited to the BindingDB/ChEMBL compound collection.

Target family distribution

We then studied the protein target family distribution of all activity cliffs. The results for the K_i/IC_{50} - and K_i -based distributions are reported in Table 2A and 2B, respectively. For the top 10 families in Table 2A,B, ranked according to the total number of activity cliffs, significant differences in cliff numbers were observed. However, on a relative scale (with respect to the total number of pairs of similar compounds), activity cliffs were similarly distributed over these target families. The top 10 families included popular therapeutic

Table 1: Activity cliff statistics

Potency annotations	Cliff degree							Cliff directionality			All cliffs
	1	2	3	4	5	6	7	dir.	undir.	Poly-cliffs	
K_i/IC_{50} values	36 063	1654	233	43	29	21	2	38 019	26	79	38 045
K_i values	10 063	330	61	17	2	0	0	10 469	4	4	10 473

Activity cliff statistics are reported for the K_i/IC_{50} - and K_i -based analyses. 'Cliff degree' denotes the number of targets per activity cliff. 'All cliffs' gives the total number of single- and multitarget cliffs. Under 'Cliff directionality' (see text), the number of directed ('dir.') and undirected ('undir.') multitarget cliffs is reported. In addition, the number of polypharmacological cliffs ('Poly-cliffs'; see text) is given.

Table 2: Target family distribution of activity cliffs

(A)

Target family	Cliff degree							Multitarget cliffs (%)	All cliffs	Cliffs (%)	Targets
	1	2	3	4	5	6	7				
Short peptide receptor	6657	161	17	1	0	0	0	2.6	6836	2.6	37
Peptidase S1	4006	68	5	3	1	0	0	1.9	4083	3.4	21
Tyr protein kinase	2656	192	37	0	0	0	0	7.4	3085	2.9	29
Peptidase A1	1610	13	6	0	0	0	0	1.2	1629	2.7	69
Prostaglandin G/H synthase	1585	39	0	0	0	0	0	2.4	1624	7.6	2
AGC Ser/Thr protein kinase	927	162	71	18	22	21	0	24.1	1221	4.3	16
Peptidase M10A	984	152	32	19	4	0	2	17.5	1193	2.6	9
CMGC Ser/Thr protein kinase	1052	75	0	0	1	0	0	6.7	1128	2.5	12
Nuclear hormone receptor	899	113	0	0	0	0	0	11.2	1012	2.2	17
Nucleotide-like receptor	955	26	4	0	0	0	0	3.0	985	1.3	6

(B)

Target family	Cliff degree					Multitarget cliffs (%)	All cliffs	Cliffs (%)	Targets
	1	2	3	4	5				
Short peptide receptor	3203	75	9	0	0	2.6	3287	2.6	32
Peptidase S1	1732	18	0	3	1	1.3	1754	2.4	18
Monoamine receptor	677	39	20	2	1	8.4	739	0.9	24
Nucleotide-like receptor	630	30	0	0	0	4.5	660	0.6	5
Alpha carbonic anhydrase	533	15	0	0	0	2.7	548	3.6	7
Peptidase C1	457	58	10	0	0	13.0	525	9.1	5
G protein-coupled receptor 2	446	0	0	0	0	0.0	446	3.4	2
Lipid-like ligand receptor	406	5	0	0	0	1.2	411	1.6	12
Nuclear hormone receptor	200	8	0	0	0	3.8	208	1.6	11
TKL Ser/Thr protein kinase	198	0	0	0	0	0.0	198	28.5	1

In (A) and (B), the target family distribution of activity cliffs is reported for the K_i/IC_{50} - and K_i -based analyses, respectively. In each case, the top 10 target families are ranked according to the number of single- and multitarget cliffs they cover. The percentage of multitarget cliffs among all activity cliffs is also reported. Furthermore, for each family, the number of compound pairs that form activity cliffs divided by the number of qualifying pairs of similar compounds and the number of targets for which activity cliffs occur are reported in the columns 'Cliffs (%)' and 'Targets', respectively.

targets for which many qualifying active compound pairs were available. Most activity cliffs were found for ligands of the short peptide receptor and peptidase S1 families. The family rankings differed for the K_i/IC_{50} -based and K_i -based cliff distributions, but in both cases, protease, kinase, nuclear hormone receptor, and G-protein-coupled receptor families were found among the top 10 families. For most highly ranked families, the percentage of multitarget cliffs among all activity cliffs was small (i.e., 0% to less than 10%), although there were exceptions; for example, for the AGC Ser/Thr kinase and the peptidase M10A (matrix metallo proteases) families,

24.1% and 17.5% of all activity cliffs were multitarget cliffs, respectively (Table 2A), and for the peptidase C1 family, 13.0% were multitarget cliffs (Table 2B).

Activity cliff directionality

Next we analyzed the directionality of multitarget activity cliffs. In a 'directed' multitarget cliff pair, the potency of compound **A** is consistently high for all targets and the potency of compound **B** is consistently low. By contrast, in an 'undirected' multitarget cliff pair,

compound **A** has high potency for at least one target for which compound **B** is only weakly potent and *vice versa*. Differences in cliff directionality are illustrated in Figure 1C,D where an exemplary directed and undirected triple-target cliff is shown, respectively. Importantly, only undirected multitarget activity cliffs contain compounds with different target selectivity, whereas in a directed cliff, the same compound is highly potent against all targets and thus always selective. In Table 1, we also report the number of undirected multitarget activity cliffs for both cliff distributions. In the K_i/IC_{50} -based distribution, only 26 of 1982 multitarget cliffs (approximately 1.3%) were undirected, and in the K_i -based distribution, only 4/410 were undirected (approximately 1.0%). Thus, nearly all multitarget cliffs were directed and activity cliff-forming compounds with different target selectivity were rare, which we considered a rather unexpected finding.

Polypharmacological cliffs

In addition, we also searched for what we regard as 'polypharmacological cliffs', i.e., multitarget activity cliffs that involved targets belonging to different protein families. Here, only targets with unambiguous family assignments (see Materials and Methods) were considered. For the K_i/IC_{50} -based distribution, we found 79 polypharmacological cliffs that involved a total of 84 compounds. Seventy-one of these cliffs were dual- and eight triple-target cliffs. For the K_i -based distribution, we identified only four (dual target) polypharmacological cliffs involving seven compounds. Hence, compounds with activity against different target families displayed only limited polypharmacological cliff potential.

Conclusions

We have carried out a comprehensive analysis of activity cliffs formed by currently available bioactive compounds. We have searched for cliffs of large magnitude that are in general least affected by measurement inaccuracies and that usually provide focal points of SAR exploration. Furthermore, in our analysis, we have differentiated between single- and multitarget activity cliffs, studied the directionality of multitarget cliffs, and also introduced polypharmacological cliffs, a special type of multitarget cliff. In general, single-target cliffs occurred much more frequently than multitarget cliffs and were similarly distributed over different target families. We also found that compounds having different target selectivity only rarely occurred in multitarget cliffs. Although the percentage of qualifying compound pairs that formed activity cliffs was only approximately 2%, on average, more than 10% of compounds active against different target families were involved in the formation of large-magnitude activity cliffs. Thus, for an active compound of interest, a thorough search of its structural neighborhood is rather likely to reveal activity cliffs from which SAR determinants might be deduced.

Acknowledgment

The authors thank Mathias Wawer for helpful discussions.

References

1. Bajorath J., Peltason L., Wawer M., Guha R., Lajiness M.S., Van Drie J.H. (2009) Navigating structure-activity landscapes. *Drug Discov Today*;14:698–705.
2. Wassermann A.M., Wawer M., Bajorath J. (2010) Activity landscape representations for structure-activity relationship analysis. *J Med Chem*;53:8209–8223.
3. Maggiora G.M. (2006) On outliers and activity cliffs – why QSAR often disappoints. *J Chem Inf Model*;46:1535.
4. The emerging concepts of activity landscapes and activity cliffs and their role in drug research. Symposium at the Fall 2010 National Meeting of the American Chemical Society, August 22–26, 2010; Maggiora, G., Lajiness, M., Bajorath, J.; Organizers.
5. Medina-Franco J.L., Martínez-Mayorga K., Bender A., Marín R.M., Giulianotti M.A., Pinilla C., Houghten R.A. (2009) Characterization of activity landscapes using 2D and 3D similarity methods: consensus activity cliffs. *J Chem Inf Model*;49:477–491.
6. Peltason L., Iyer P., Bajorath J. (2010) Rationalizing three-dimensional activity landscapes and the influence of molecular representations on landscape topology and the formation of activity cliffs. *J Chem Inf Model*;50:1021–1033.
7. Dimova D., Wawer M., Wassermann A.M., Bajorath J. (2011) Design of multi-target activity landscapes that capture hierarchical activity cliff distributions. *J Chem Inf Model*;51:258–266.
8. Paolini G.V., Shapland R.H.B., van Hoorn W.P., Mason J.S., Hopkins A.L. (2006) Global mapping of pharmacological space. *Nat Biotechnol*;24:805–815.
9. Bajorath J. (2008) Computational analysis of ligand relationships within target families. *Curr Opin Chem Biol*;12:352–358.
10. Mestres J., Gregori-Puigjané E. (2009) Conciliating binding efficiency and polypharmacology. *Trends Pharmacol Sci*;30:470–474.
11. Liu T., Lin Y., Wen X., Jorissen R.N., Gilson M.K. (2007) Binding-DB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res*;35:D198–D201.
12. Overington J. (2009) ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr. *J Comput Aided Mol Des*;23:195–198.
13. Rogers D., Hahn M. (2010) Extended-connectivity fingerprints. *J Chem Inf Model*;50:742–754.
14. Wawer M., Bajorath J. (2010) Similarity-potency trees: a method to search for SAR information in compound data sets and derive SAR rules. *J Chem Inf Model*;50:1395–1409.
15. UniProt Consortium. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res*;38:D142–D148.

Notes

^aPubchem, <http://pubchem.ncbi.nlm.nih.gov>.

^bScitegic Pipeline Pilot, Accelrys Inc., San Diego, CA, USA, <http://accelrys.com/products/scitegic/index.html>.

^cMACCS Structural Keys, Symyx Technologies, Inc., Sunnyvale, CA, USA, <http://www.symyx.com>.

Summary

Herein, the distribution and directionality of activity cliffs and the propensity of currently available bioactive compounds to form activity cliffs were systematically determined. Single-target cliffs were most frequently observed accounting for 94.8% of all cliffs. Surprisingly, these cliffs were evenly distributed over different target families. Furthermore, the majority of the multi-target cliffs were uni-directional and of low degree (i.e., number of targets involved), yet instances of multi-target cliffs of up to degree 7 were also detected. Only a limited number of “polypharmacological” cliffs was identified. In addition, approximately 12% of all bioactive compounds were involved in the formation of at least one activity cliff. My contribution to the study reported herein has been to aid in the assessment and large-scale analysis of currently available activity cliffs.

Activity cliffs are considered centers of SAR discontinuity and provide direct access to SAR information. Although it is well-appreciated that different molecular representations will inevitably affect SAR analysis, it is currently unknown to what extent the use of alternative representation might change the nature of SARs. To shed light on this question, we have carried out a large scale analysis of fingerprint dependent changes in SAR information.

Chapter 5

Quantifying the Fingerprint Descriptor Dependence of Structure-Activity Relationship Information on a Large Scale

Introduction

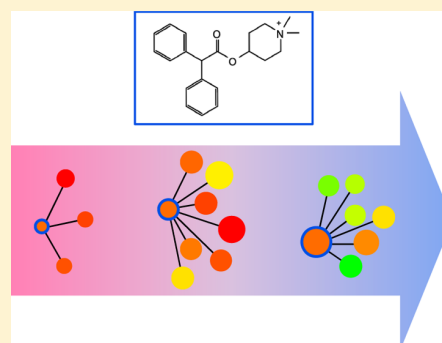
It is well-understood that the quantitative assessment of molecular similarity is highly dependent on the chosen molecular representation. Structural similarity and potency comparisons between bioactive compounds are integral part of SAR analysis and hence the outcome of the SAR study will inevitably change when alternative representations are used. Numerical analysis functions (e.g., SARI, SALI) have been introduced that can be used as a diagnostic of global and local SAR characteristics. Herein, we address the question of how to quantitatively assess the fingerprint dependence of SAR information. By systematically calculating local per-compound discontinuity scores, changes in the SAR phenotype for individual compounds can be monitored and fingerprint-dependent changes identified.

Quantifying the Fingerprint Descriptor Dependence of Structure–Activity Relationship Information on a Large Scale

Dilyana Dimova,[†] Dagmar Stumpfe,[†] and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113 Bonn, Germany

ABSTRACT: It is well-known that different molecular representations, e.g., graphs, numerical descriptors, fingerprints, or 3D models, change the numerical results of molecular similarity calculations. Because the assessment of structure–activity relationships (SARs) requires similarity and potency comparisons of active compounds, this representation dependence inevitably also affects SAR analysis. But to what extent? How exactly does SAR information change when alternative fingerprints are used as descriptors? What is the proportion of active compounds with substantial changes in SAR information induced by different fingerprints? To provide answers to these questions, we have quantified changes in SAR information across many different compound classes using six different fingerprints. SAR profiling was carried out on 128 target-based data sets comprising more than 60 000 compounds with high-confidence activity annotations. A numerical measure of SAR discontinuity was applied to assess SAR information on a per compound basis. For ~70% of all test compounds, changes in SAR characteristics were detected when different fingerprints were used as molecular representations. Moreover, the SAR phenotype of ~30% of the compounds changed, and distinct fingerprint-dependent local SAR environments were detected. The fingerprints we compared were found to generate SAR models that were essentially not comparable. Atom environment and pharmacophore fingerprints produced the largest differences in compound-associated SAR information. Taken together, the results of our systematic analysis reveal larger fingerprint-dependent changes in compound-associated SAR information than would have been anticipated.



INTRODUCTION

The assessment of structure–activity relationship (SAR) information associated with active compounds is of central relevance in medicinal chemistry.¹ A variety of computational methods are applied to aid in SAR analysis² including QSAR approaches,³ visualization techniques,⁴ and numerical analysis functions for SAR profiling.^{5,6} SAR analysis generally requires the comparison of structures of active compounds and their potency values. Structure comparison focuses on the assessment of molecular similarity/dissimilarity, in qualitative and/or quantitative terms.^{7,8} Both from a medicinal chemistry^{1,7} and computational^{8,9} perspective, molecular similarity is often difficult to assess.

A quantitative computational assessment of similarity requires the choice of a consistently applied molecular descriptors (representations) and a similarity or distance metric/measure.^{10,11} Comparisons are typically carried out in a pairwise manner, either by quantifying the similarity of given molecular representations or by calculating the distance in chemical reference space. The molecular descriptor and, to a lesser extent, metric dependence of similarity calculations are well-appreciated caveats of molecular similarity analysis.^{8–12} Similarity relationships display a tendency to change when alternative molecular representations are used, which inevitably affects similarity search calculations¹² and quantitative SAR analysis.¹⁴ Although this dependence is widely appreciated, its potential magnitude remains unclear.

In addition to QSAR approaches focusing on individual compound series, numerical SAR analysis functions,^{5,6} such as the SAR index (SARI),⁵ are available for quantitative SAR exploration. SARI integrates similarity and potency comparisons of active compounds, thereby producing a numerical score that quantitatively characterizes SAR features. This characterization can be carried out both at the level of complete data sets (i.e., describing global SAR features) or compound subsets (local SARs). Different SAR feature categories can be distinguished on the basis of numerical SAR analysis. For example, SAR continuity refers to the presence of structurally similar compounds having comparable potency. The presence of continuous SARs provides the basis of scaffold hopping¹³ in similarity searching and virtual screening.^{9–12} By contrast, SAR discontinuity refers to the presence of structurally similar compounds with a large potency differences.^{5,14} The extreme form of SAR discontinuity is represented by activity cliffs^{14–17} that are rationalized as pairs of similar compounds (structural analogs) with large potency differences. For the analysis of activity cliffs, the definition of similarity and potency difference threshold values is typically required.¹⁶ For large-scale SAR analysis using numerical analysis functions, different types of fingerprints,^{9–12} i.e., bit representations of molecular structure and properties, are

Received: July 11, 2013

Published: August 22, 2013

generally preferred molecular representations and most widely used for similarity calculations.^{5,6,14}

It is currently unknown to what extent active compounds across different targets might change their SAR phenotype, e.g., from a continuous to a discontinuous one and vice versa, when different fingerprints are used as molecular representations. Obtaining this information would provide a first quantification of fingerprint-dependent changes in SAR characteristics on a per compound basis and help to estimate the influence of these effects on the reliability of computational SAR models. Therefore, we have systematically analyzed 128 compound sets with high-confidence activity data using different fingerprints to quantitatively describe and compare changes in SAR information at the level of individual compounds.

MATERIALS AND METHODS

Compound Data Sets. Data sets were assembled from ChEMBL (version 15).¹⁸ For our analysis, only compounds with precisely specified equilibrium constants (K_i values) below 1 μ M for human targets at the highest confidence level (confidence score 9) were selected. A compound with multiple measurements for the same target was only considered if it had consistent potency measurements (i.e., if all values fell within 1 order of magnitude). In this case, the average potency was calculated as the final annotation. In addition, each data set was required to contain at least 100 compounds. On the basis of these selection criteria, 128 target-based data sets comprising a total of 60 248 compounds were obtained. The data sets are made freely available via the downloads section of following: <http://www.lifescienceinformatics.uni-bonn.de>.

Fingerprints. Six fingerprints of different design and complexity were used as molecular representations:

- Extended connectivity fingerprint with bond diameter 4 (ECFP4)¹⁹ producing $\sim 4 \times 10^9$ theoretically possible features. ECFP4 is a topological fingerprint that encodes layered atom environments with a maximum diameter of four bonds around each atom.
- Functional class fingerprint with bond diameter 4 (FCFP4) is a closely related derivative of ECFP4 that replaces atom types with pharmacophore features (such as hydrogen-bond donors and acceptors, charged, or aromatic atoms), which renders atom information less specific compared to ECFP4 and reduces the size of feature sets.
- Molecular access system (MACCS)²⁰ consists of 166 structural fragments with 1–10 nonhydrogen atoms.
- Typed graph distances (TGD)²¹ is an atom pair-type fingerprint consisting of 420 bits. Shortest distances in the molecular graph between two atoms (represented as seven pharmacophore features) are calculated and assigned to 15 distance ranges.
- Typed graph triangles (TGT)²¹ contains 1704 bits positions representing three-point pharmacophore patterns in molecular graphs. Each atom is assigned to one of four atom types (hydrogen-bond donor and acceptor, donor/acceptor, or hydrophobic), and six bond distance ranges are applied.
- Donor–acceptor polar-hydrophobe triangle (Gpi-DAPH3)²¹ is a molecular graph-based three-point pharmacophore fingerprint generating 30 240 possible features. In this case, each atom is assigned to one of eight atom types derived from three atomic properties (π

system, donor, or acceptor), and eight bond distance ranges are utilized.

All fingerprint representations were calculated using the Molecular Operating Environment (MOE).²¹ The TGD, TGT, and GpiDAPH3 fingerprints are MOE-internal developments.

SARI Discontinuity Score. SARI is a numerical SAR analysis function to quantify SAR features of compound data sets and consists of separate terms accounting for SAR continuity and discontinuity, respectively.⁵ We have used the discontinuity score component as a quantitative measure of per compound SAR discontinuity. The raw (non-normalized) discontinuity score is defined as⁵

$$\text{raw}_{\text{disc}} = \frac{\sum_{\{i,j|\text{sim}(i,j)>x,i>j\}} |\text{pot}(i) - \text{pot}(j)| \cdot \text{sim}(i,j)}{|\{i,j|\text{sim}(i,j) > x, i > j\}|}$$

Here, $\text{pot}(i)$ is the potency ($\text{p}K_i$) value of compound i and $\text{sim}(i,j)$ the calculated fingerprint similarity for compounds i and j . The raw discontinuity score is derived from the average of potency differences between pairs of ligands multiplied by their similarity. Accordingly, the score emphasizes the presence of structurally similar compounds with large potency differences. Therefore, it is advisable to limit discontinuity scoring to compounds that have at least limited molecular similarity.⁵ Hence, a similarity threshold value (x) is usually applied, as further discussed below.

Depending on the normalization procedure, raw discontinuity scores can be converted into scores that either quantify the degree of SAR discontinuity for a complete data set, compound subsets, or individual compounds.^{5,22,23} For our analysis, we utilized per compound discontinuity scores to assess local SAR discontinuity. Therefore, for each fingerprint, raw discontinuity scores were systematically calculated for all 128 classes. The raw scores were then converted into Z-scores using the sample mean and standard deviation of the score distribution for each individual data set, i.e., the intraclass score distribution for each fingerprint. Finally, cumulative probabilities were calculated to map scores onto the value range [0, 1].⁵ All similar compound pairs were taken into consideration, regardless of potency differences, to fully account for the structural neighborhood of each individual compound,^{22,23} a prerequisite for assessing local SAR discontinuity. As shown herein, per compound discontinuity scores provide a meaningful measure of compound-associated SAR information and make it possible to characterize local SAR environments in compound data sets in detail.

Similarity Calculation and Threshold Correspondence. Pairwise fingerprint similarity was quantified using the Tanimoto coefficient (T_c).¹¹ As a reference point for the discontinuity score similarity threshold, a MACCS T_c of 0.70 was applied, which typically indicates remote similarity.¹² Compound pairs yielding further increasing MACCS T_c values become increasingly similar (as an activity cliff criterion, a MACCS T_c of 0.85 is often applied).¹⁶ Across all 128 data sets, a MACCS T_c of 0.70 yielded 12% of all possible compound pairs meeting or exceeding this threshold.

In order to determine corresponding T_c threshold values for all fingerprints yielding the same number of similar compound pairs, a three-step procedure was applied: (a) For each of the 128 data sets, the number of compound pairs yielding a MACCS $T_c \geq 0.70$ was determined; (b) for each of the remaining fingerprints, the T_c threshold was determined that yielded the same number of compound pairs for each set; and

(c) for each fingerprint, the median of its 128 individual threshold values was calculated and used as its global threshold.

On the basis of this analysis, the following corresponding T_c threshold values were derived: MACCS, 0.70; ECFP4, 0.31; FCFP4, 0.38; TGD, 0.69; TGT, 0.65; GpiDAPH3, 0.15. These thresholds were applied for the calculation of per-compound discontinuity scores.

RESULTS AND DISCUSSION

Study Concept and Goals. We have aimed to quantify the influence of different fingerprint descriptors on compound-

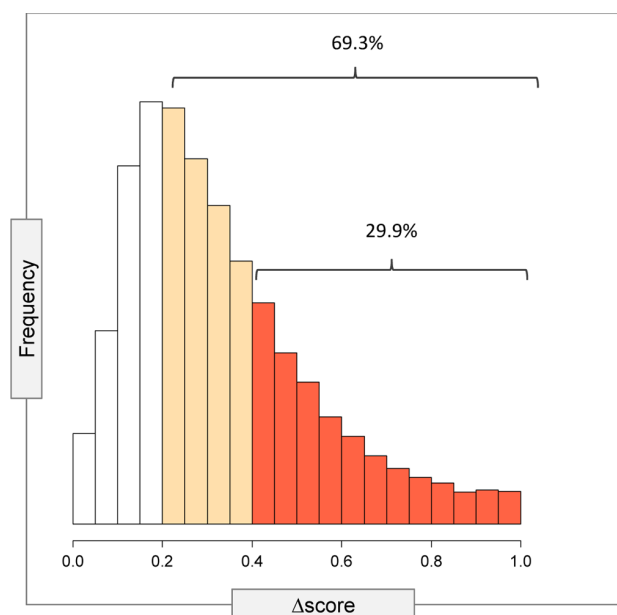


Figure 1. Distribution of pairwise score differences. For each compound, discontinuity score differences are calculated for all pairs of fingerprints, yielding a total of 15 score differences for an individual compound. The histogram reports the distribution of the maximal score differences for all 60 248 active compounds.

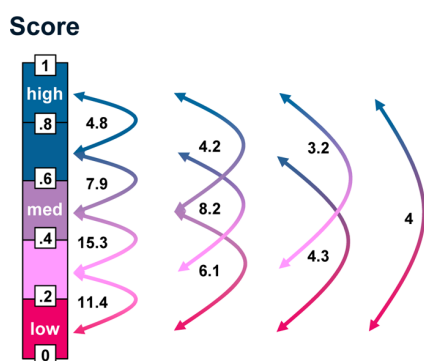


Figure 2. Discontinuity score shifts. Five score ranges ($[0, 0.2]$, $(0.2, 0.4]$, $(0.4, 0.6]$, $(0.6, 0.8]$, $(0.8, 1]$) are defined. Ranges of low, medium, and high discontinuity are marked. Arrows represent largest observed score differences and are labeled with the percentage of compounds displaying such fingerprint-dependent differences.

associated SAR information on a large scale across many different targets. To our knowledge, this is the first study that systematically and quantitatively assesses fingerprint-dependent

Table 1. Distribution of Score Differences for Fingerprint Combinations^a

fingerprint combination	$\Delta\text{score} \geq 0.2$	$\Delta\text{score} \geq 0.4$
ECFP4/FCFP4	12.2	3.8
ECFP4/GpiDAPH3	20.5	6.9
ECFP4/MACCS	20.0	6.9
ECFP4/TGT	28.6	11.2
ECFP4/TGD	31.1	12.7
FCFP4/GpiDAPH3	20.1	7.0
FCFP4/MACCS	20.7	7.4
FCFP4/TGT	24.6	9.7
FCFP4/TGD	26.7	10.9
GpiDAPH3/MACCS	25.6	9.1
GpiDAPH3/TGT	25.8	9.9
GpiDAPH3/TGD	27.0	10.7
MACCS/TGT	28.9	11.7
MACCS/TGD	30.2	12.8
TGT/TGD	18.8	7.9

^aFor all possible combinations of fingerprints, the percentage of compounds with discontinuity score differences (Δscore) of at least 0.2 or 0.4 is reported as an average over all classes.

Table 2. Number of Fingerprint Combinations Producing Large Score Changes^a

no. FP combinations	$\Delta\text{score} \geq 0.2$			$\Delta\text{score} \geq 0.4$		
	no. cpds	% cpds	avg. no. neigh	no. cpds	% cpds	avg. no. neigh
0	18 509	30.7	108.2	42 217	70.1	94.1
1	4155	6.9	96.6	3639	6.0	66.5
2	4079	6.8	88.5	3045	5.1	60.9
3	3655	6.1	82.3	2358	3.9	55.1
4	4077	6.8	78.9	2255	3.7	54.0
5	7126	11.8	64.3	2822	4.7	38.4
6	3459	5.7	70.1	1044	1.7	49.1
7	3249	5.4	64.9	795	1.3	41.7
8	4774	7.9	58.2	1445	2.4	30.0
9	3552	5.9	49.7	561	0.9	24.9
10	1671	2.8	48.6	52	0.1	19.4
11	1510	2.5	37.5	15	0.0	21.3
12	360	0.6	32.2	0	0	0
13	71	0.1	23.6			not possible
14	1	0.0	3.2			not possible
15	0	0	0			not possible

^aThe number of fingerprint (FP) combinations yielding a score difference (Δscore) of at least 0.2 and 0.4 is determined for all compounds. For all combinations, the corresponding number and percentage of compounds (cpds) are reported. In addition, the average number of structural neighbors (no. neigh) calculated for all fingerprints (no. FP = combinations 0) and for only those fingerprints that participate in pairs with $\Delta\text{score} \geq 0.2$ or $\Delta\text{score} \geq 0.4$ (no. FP combinations ≥ 1) are reported. It should be noted that the maximal number of fingerprint combinations that can yield $\Delta\text{score} \geq 0.2$ and $\Delta\text{score} \geq 0.4$ is 15 and 12, respectively. Hence, it is not possible that all numbers of combinations of six fingerprints can meet the $\Delta\text{score} \geq 0.4$ condition. The three numbers of combinations for which $\Delta\text{score} \geq 0.4$ cannot be obtained (13, 14, and 15) are designated as "not possible".

changes in SAR information content at the level of individual active compounds. To these ends, we have selected more than 60 000 compounds with high-confidence activity annotations for 128 targets and calculated pairwise similarities using six fingerprints of different design and complexity. For these

Table 3. Compound Neighborhood Statistics for Different Fingerprint Combinations^a

no. FP combinations	Δ score ≥ 0.2		Δ score ≥ 0.4	
	category A	category B	category A	category B
0	2.9	97.1	2.3	97.6
1	1.1	98.9	4.4	95.5
2	1.6	98.4	6.1	93.9
3	2.1	97.9	8.2	91.8
4	3	97.0	10.4	89.6
5	9.9	90.2	29	70.9
6	4.2	95.8	15.8	84.2
7	4.6	95.4	23.2	76.8
8	11.5	88.5	39.2	60.7
9	14.3	85.7	46.5	53.4
10	13.5	86.5	62.4	37.7
11	25.3	74.7	69.1	30.9
12	30.7	69.3	0	0
13	42.4	57.6	not possible	
14	78.5	21.4	not possible	
15	0	0	not possible	

^aThe number of fingerprint combinations yielding a score difference (Δ score) of at least 0.2 and 0.4 is reported for all compounds. For each pair of fingerprints, the fingerprint-dependent compound neighborhoods are determined and classified as distinct (i.e., two fingerprints produce different sets of structural neighbors; category A) or overlapping (i.e., two fingerprints produce a subset of shared structural neighbors; category B). In addition, the average percentage of distinct and overlapping neighborhoods calculated for all fingerprints (no. FP combinations = 0) and for only those fingerprints that participate in pairs yielding Δ score ≥ 0.2 or Δ score ≥ 0.4 (no. FP combinations ≥ 1) are reported. Three numbers of fingerprint combinations for which Δ score ≥ 0.4 cannot be obtained (see legend of Table 2) are designated as “not possible”.

fingerprints, corresponding T_c threshold values were determined to ensure that the same proportion of all possible compound pairs was classified as similar.

SAR discontinuity is an indicator of SAR information content.^{14,15} In order to account for SAR discontinuity in a consistent manner, a SAR discontinuity scoring scheme was applied that quantifies the discontinuity contributions of individual compounds.¹⁴ The latter criterion is of critical relevance for our current analysis. What does local SAR discontinuity mean? Local SAR discontinuity is high if a compound has a potency value that substantially deviates from the potency of its structural neighbors.

Through systematic assessment of compound-centric SAR discontinuity, fingerprint-dependent changes in SAR information were quantified for all test compounds and local SAR environments were analyzed.

Distribution of Discontinuity Score Differences. For each compound, discontinuity scores using all 6 fingerprints and score differences for all possible 15 pairwise fingerprint combinations were calculated. In Figure 1, the distribution of the maximal score differences per compound is reported. For $\sim 70\%$ and $\sim 30\%$ of all compounds, maximal score differences (Δ score) ≥ 0.2 and ≥ 0.4 were detected, respectively. Discontinuity score differences ≥ 0.2 are substantial, and score differences ≥ 0.4 indicate a major change in SAR information associated with a given compound. This is the case because a score change of magnitude 0.4 or larger transforms a compound with low SAR discontinuity into one with intermediate or high discontinuity and vice versa, as further

discussed below. Thus, for $\sim 30\%$ of all active compounds across the 128 target-based data sets, large-magnitude changes in SAR information were observed that altered their SAR phenotype, giving rise to fingerprint-dependent SAR phenotype switches, which were quantified for the first time. Overall, an unexpectedly high proportion of compounds active against 128 different targets was found to change their SAR phenotype when different fingerprints were used. On the basis of fingerprint similarity value distributions,²⁴ a much smaller proportion of compounds would have been predicted. This meant that the structural neighborhood of many compounds substantially changed when different fingerprints were used, thereby completely changing many local SAR environments, as further discussed below. In Figure 2, the range distribution of score differences is reported. Differences were not confined to certain score subranges. Rather, shifts of varying magnitudes across the entire discontinuity score range were observed.

Fingerprint Comparison. In Table 1, the proportion of compounds with discontinuity score differences ≥ 0.2 or ≥ 0.4 is reported for all 15 pairwise fingerprint combinations, ranging from 12.2–31.1% (score difference ≥ 0.2) and 3.8–12.8% (≥ 0.4). All fingerprint combinations were found to induce substantial discontinuity score differences including closely related fingerprints, such as ECFP4/FCFP4, for which 12.2% and 3.8% of the compounds displayed score differences ≥ 0.2 and ≥ 0.4 , respectively. The comparison made it possible to identify combinations of fingerprint types that generated inconsistent SAR models. A major finding has been that atom environment fingerprints and 2D pharmacophore fingerprints produced largely different SAR models. For example, the ECFP4/TGD combination yielded score differences ≥ 0.2 and ≥ 0.4 for 31.1% and 12.7% of active compounds, respectively. On the basis of the fingerprint pair-based score differences in Table 1, all 15 fingerprint combinations were considered for further analysis of discontinuity score distributions.

Score Differences and Fingerprint Combinations. Table 2 reports the number of fingerprint combinations that produced large score differences and the corresponding numbers of compounds. In addition, the average number of neighbors (meeting or exceeding the fingerprint similarity thresholds) of these compounds is reported. For 30.7% and 70.1% of all compounds, no fingerprint combination generated score differences ≥ 0.2 and ≥ 0.4 , respectively. These data are consistent with the score difference distribution in Figure 1. However, 1–11 different fingerprint combinations yielded score differences ≥ 0.2 for thousands of compounds each and 1–8 combinations differences ≥ 0.4 for comparable numbers of compounds. Thus, multiple fingerprint combinations were generally responsible for substantial changes in SAR discontinuity, and large proportions of active compounds experienced large changes in their structural neighborhoods for multiple fingerprints.

Fingerprint Combinations and Compound Neighborhoods. Table 2 also reveals a steady decrease in the number of structural neighbors of compounds for which increasing numbers of fingerprint combinations generated large changes in SAR discontinuity. For compounds that did not yield score differences ≥ 0.2 and ≥ 0.4 , on average more than 100 and 90 structural neighbors were detected, respectively. Thus, these compounds were very similar to many others. By contrast, for compounds for which increasing numbers of fingerprint combinations led to substantial score changes, decreasing numbers of structural neighbors were detected. For example,

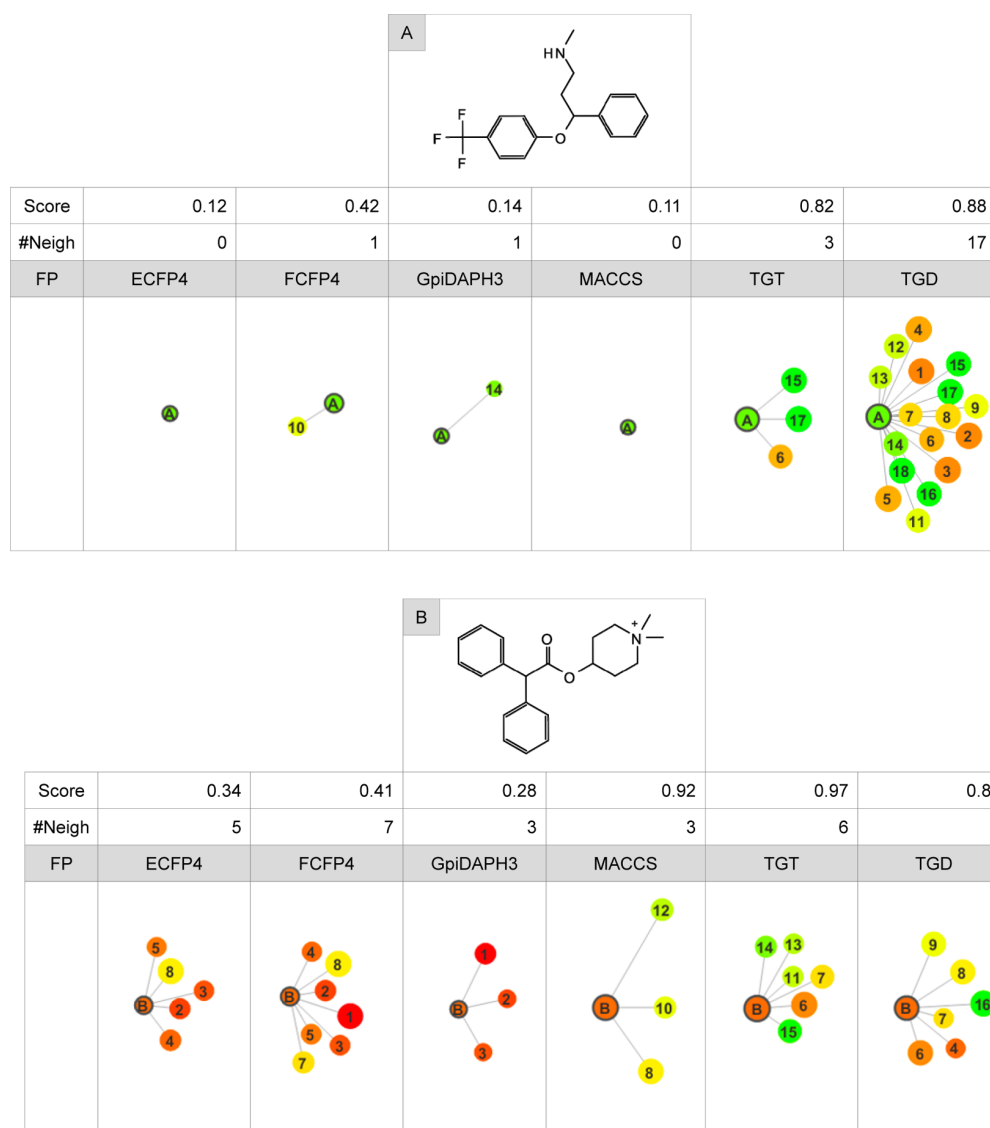


Figure 3. Local SAR environments. In (A) and (B), the chemical neighborhood of two exemplary compounds in their data sets is shown when similarity relationships are calculated using the six different fingerprints. Compounds A and B are muscarinic acetylcholine receptor M5 antagonists (ChEMBL ID 2035). Sections of network-like similarity graphs (NSGs)²³ are displayed representing the fingerprint-dependent neighborhoods of each compound. In NSGs, compounds are represented as nodes and edges similarity relationships. For each fingerprint, the corresponding Tc similarity threshold values reported in the Materials and Methods section are applied as a similarity (edge) criterion. Nodes are numbered, color-coded by compound potency using a color spectrum ranging from red (high) over yellow (intermediate) to green (low potency) and scaled in size by local discontinuity scores. For compounds A and B and each fingerprint, the discontinuity score and the number of structural neighbors (#Neigh) applying the fingerprint-dependent Tc threshold values are reported.

for compounds whose discontinuity scores were notably changed by eight different fingerprint combinations, on average only 58 ($\Delta\text{score} \geq 0.2$) and 30 ($\Delta\text{score} \geq 0.4$) neighbors were identified (Table 2). Thus, compounds whose SAR characteristics were increasingly fingerprint-dependent formed fewer similarity relationships than other compounds; another unexpected finding.

Compound neighborhoods are further analyzed in Table 3. Most of the compound neighborhoods generated with different fingerprints were overlapping (Category B). However, for compounds for which multiple fingerprint combinations produced substantial score changes, increasing numbers of distinct neighborhoods (Category A) were observed, i.e., neighborhoods that did not share any compounds. For

example, for compounds whose scores were significantly altered by eight different fingerprint combinations, on average $\sim 12\%$ ($\Delta\text{score} \geq 0.2$) and $\sim 39\%$ ($\Delta\text{score} \geq 0.4$) of their neighborhoods generated with these fingerprints were distinct (Table 3). Thus, different fingerprints often produced nonoverlapping similarity relationships that led to large changes in local SAR discontinuity.

Local SAR Environments. In Figure 3, local SAR environments centered on exemplary compounds are compared for different fingerprints with the aid of similarity-based molecular networks,²³ which illustrate changes in local SAR information. In Figure 3A, an active compound is shown for which two fingerprints (ECFP4, MACCS) did not yield structural neighbors. Thus, in these cases, the compound was

a singleton carrying essentially no SAR information. In addition, two other fingerprints (FCFP4, GpiDAPH3) identified only a single neighbor. By contrast, TGT and especially TGD detected multiple neighbors with varying potency. Thus, in these local environments, the test compound introduced a high degree of SAR discontinuity. Discontinuity scores for this compound varied from 0.11 (MACCS) to 0.88 (TGD). In the TGD-dependent neighborhood, the compound formed multiple activity cliffs (combinations of large red and green nodes in Figure 3A) and hence obtained a high discontinuity score. Thus, the SAR character of this compound and its associated SAR information content completely changed dependent on the fingerprint that was used.

Furthermore, in Figure 3B, another exemplary compound is shown for which different fingerprints consistently detected multiple neighbors. The fingerprint-dependent neighborhoods were partly overlapping and distinct and resulted in very different local SAR environments, reflected by either low or high discontinuity scores ranging from 0.28 (MACCS) to 0.97 (TGT). For MACCS, TGD, and TGT, the test compound formed increasing numbers of activity cliffs in its neighborhood and was thus highly discontinuous. Thus, in these cases, this compound would be assigned a key compound for SAR analysis. By contrast, in the ECFP4-, FCFP4-, and GpiDAPH3-dependent environments, most structural neighbors had potency values very similar to the test compound, which resulted in low/intermediate discontinuity scores. In these cases, the compound would not be considered as a focal point of SAR analysis. Thus, the SAR characteristic of the compound in Figure 3B also fundamentally changed with different fingerprint representations.

The exemplary compounds in Figure 3 had score differences ≥ 0.4 for eight (compound A) and nine (B) different fingerprint combinations, respectively. Score differences were in seven (A) and four (B) cases due to the presence of distinct neighbors and in one (A) and five (B) cases due to partly overlapping fingerprint-dependent neighborhoods. Thus, the local SAR environments of these exemplary compounds often fundamentally differed. These findings illustrate that a very different conclusion concerning local SAR features in compound data sets might be drawn when alternative fingerprints are used.

CONCLUSIONS

Herein we have reported a first quantitative and large-scale analysis of fingerprint-dependent changes in SAR information associated with $\sim 60\,000$ compounds active against 128 different targets. Different types of 2D fingerprints were compared including structural fragment, atom environment, and pharmacophore fingerprints. The analysis was facilitated by a consistent quantitative assessment of per compound SAR discontinuity and revealed a number of unexpected findings, as summarized in the following: For $\sim 70\%$ of all compounds, substantial changes in SAR discontinuity were detected, and for $\sim 30\%$ of all compounds, different numbers of fingerprint combinations were found to change their SAR phenotype; a much larger proportion of compounds that would have been estimated on the basis of similarity score distributions. Furthermore, we have found that compounds forming below-average numbers of similarity relationships to others displayed a particularly strong fingerprint dependence of their local SAR environments. Moreover, we have determined that in nearly all of the combinations of fingerprints we analyzed (except the most closely related ones such as ECFP4 and FCFP4), we led

to large differences in compound-associated SAR information, more so than we had anticipated. Atom environment and pharmacophore fingerprints produced largest differences in SAR information. A major conclusion from our study is that SAR phenotypes of many compounds change when alternative fingerprints are used. Thus, ensuing SAR differences are not gradual but often fundamental. This makes it essentially impossible to draw general conclusions concerning local SARs and individual compounds. Key compounds representing centers of SAR discontinuity in data sets identified with a given fingerprint might not be detected when other fingerprints are used. In SAR analysis, awareness should be raised concerning the unexpectedly large magnitude of these effects and the resulting inconsistency of many SAR models produced by widely used fingerprints. In the future, the current analysis might be further extended to alternative sets of numerical descriptors, which are typically preferred for QSAR modeling.

AUTHOR INFORMATION

Corresponding Author

*E-mail: bajorath@bit.uni-bonn.de

Author Contributions

[†]These authors contributed equally.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

D.S. is supported by Sonderforschungsbereich 704 of the Deutsche Forschungsgemeinschaft.

REFERENCES

- (1) Wermuth, C. G. *The Practice of Medicinal Chemistry*, 3rd ed.; Academic Press: Waltham, MA, 2008.
- (2) Wawer, M.; Louunkine, E.; Wassermann, A. M.; Bajorath, J. Data Structures and Computational Tools for the Extraction of SAR Information from Large Compound Sets. *Drug Discovery Today* **2010**, *15*, 631–639.
- (3) Esposito, E. X.; Hopfinger, A. J.; Madura, J. D. Methods for Applying the Quantitative Structure-Activity Relationship Paradigm. *Methods Mol. Biol.* **2004**, *275*, 131–214.
- (4) Stumpfe, D.; Bajorath, J. Methods for SAR Visualization. *RSC Adv.* **2012**, *2*, 369–378.
- (5) Peltason, L.; Bajorath, J. SAR Index: Quantifying the Nature of Structure-Activity Relationships. *J. Med. Chem.* **2007**, *50*, 5571–5578.
- (6) Guha, R.; Van Drie, J. H. Structure-Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* **2008**, *48*, 646–658.
- (7) Kubinyi, H. Similarity and Dissimilarity: A Medicinal Chemist's View. *Perspect. Drug Discovery Des.* **1998**, *9–11*, 225–232.
- (8) *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; John Wiley & Sons: New York, 1990.
- (9) Eckert, H.; Bajorath, J. Molecular Similarity Analysis in Virtual Screening: Foundations, Limitations and Novel Approaches. *Drug Discovery Today* **2007**, *12*, 225–233.
- (10) Willett, P. Searching Techniques for Databases of Two- and Three-Dimensional Structures. *J. Med. Chem.* **2005**, *48*, 4183–4199.
- (11) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
- (12) Stumpfe, D.; Bajorath, J. Similarity Searching. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 260–282.
- (13) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. Scaffold-Hopping by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed.* **1999**, *19*, 2894–2896.

- (14) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity Landscape Representations for Structure–Activity Relationship Analysis. *J. Med. Chem.* **2010**, *53*, 8209–8223.
- (15) Maggiora, G. M. On Outliers and Activity Cliffs – Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535–1535.
- (16) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J. Med. Chem.* **2012**, *55*, 2932–2942.
- (17) Medina-Franco, J. L.; Martínez-Mayorga, K.; Bender, A.; Marin, 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.
- (18) 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.* **2011**, *40*, D1100–D1107.
- (19) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
- (20) MACCS Structural Keys; Accelrys: San Diego, CA.
- (21) Molecular Operating Environment (MOE); Chemical Computing Group Inc.: Montreal, Quebec, Canada.
- (22) Peltason, L.; Hu, Y.; Bajorath, J. From Structure-Activity to Structure-Selectivity Relationships: Quantitative Assessment, Selectivity Cliffs, and Key Compounds. *ChemMedChem.* **2009**, *4*, 1864–1873.
- (23) 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. *J. Med. Chem.* **2008**, *51*, 6075–6084.
- (24) Vogt, M.; Stumpfe, D.; Geppert, H.; Bajorath, J. Scaffold Hopping Using Two-dimensional Fingerprints: True Potential, Black Magic, or a Hopeless Endeavor? Guidelines for Virtual Screening. *J. Med. Chem.* **2010**, *53*, 5707–5715.

Summary

A large-scale SAR profiling analysis was carried out on more than 60,000 compounds that were organized into 128 target-based sets. Local per-compound discontinuity scores were calculated and changes in the score for nearly 70% of the compounds observed. Furthermore, for $\sim 30\%$ of all compounds, the use of alternative fingerprint representations led to a change in the SAR phenotype. In addition, compounds involved in relatively limited numbers of structural relationships displayed a stronger tendency to change their SAR characteristics when alternative fingerprints were used. In light of these observations, SAR characteristics often considerably changed and hence were highly fingerprint-dependent. My contributions to this work have been the systematic assessment of local discontinuity scores and data analysis.

In this and the previous study, we have addressed the fingerprint-dependent changes of SAR characteristics and analyzed the frequency of occurrence of activity cliffs. Another thus far unexplored question has been whether SAR information associated with activity cliffs might more frequently result in compound series with steadily increasing potency. Therefore, we have designed a study to investigate the potential activity cliff advantage for SAR progression.

Chapter 6

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

Introduction

The SAR exploration has a major influence on compound optimization and practical medicinal chemistry. Activity cliffs are focal points of SAR analysis and have been extensively studied from many different perspectives. However, it is currently unknown if there is a detectable SAR advantage in optimizing compounds involved in activity cliffs compared to other compounds available as starting points for SAR analysis. To investigate this question, we have designed a computational compound pathway model to represent compound series with steadily increasing potency ultimately leading to highly potent data set compounds. Three major pathway categories have been introduced and distinguished by their origin. Relative pathway frequencies have been determined to

serve as a diagnostic for SAR information gain associated with different types of data set compounds.

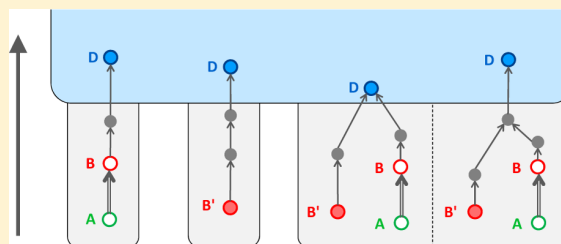
Compound Pathway Model To Capture SAR Progression: Comparison of Activity Cliff-Dependent and -Independent Pathways

Dagmar Stumpfe, Dilyana Dimova,[†] Kathrin Heikamp,[†] and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113 Bonn, Germany

Supporting Information

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



INTRODUCTION

Capturing SAR information in compound data sets of any source is a prime task in computational medicinal chemistry,^{1,2} in addition to, for example, QSAR-based compound activity predictions.³ SAR information can be extracted from compound activity data through numerical and/or graphical analysis.^{1,2} In general terms, SAR information extraction requires the systematic study of structural and potency relationships between active compounds and the identification of series with defined SAR characteristics such as SAR continuity or discontinuity.² Numerical and graphical analysis components are typically combined to model activity landscape representations of compound data sets that integrate structural and potency relationships in a systematic manner.² Cardinal features of activity landscapes are activity cliffs that are formed by pairs or groups of compounds with large potency differences.^{4–6} The activity cliff concept is popular because activity cliffs encode small structural changes leading to large potency alterations, which is generally thought to associate activity cliff compounds with high SAR information content.^{4,6} Activity landscapes and activity cliffs have been studied using a variety of molecular representations and similarity measures^{2,6–8} also including structurally conservative approaches that limit the formation of cliffs to analogous compounds.⁹ Compared to whole-molecule similarity calculations, the latter approach further supports the interpretation of activity cliffs from a chemical perspective, which is important for practical SAR analysis. Systematic surveys of activity cliffs in public domain compounds have been carried out to characterize activity cliff populations^{10,11} and their potency range distribu-

tions.¹¹ Depending on the chosen molecular representations and similarity criteria considered, only ~4–6% of all pairs of qualifying similar active compounds formed activity cliffs, confirming that cliffs are a rare activity landscape feature. On the other hand, ~20–30% of all compounds with high-confidence activity annotations across different targets were found to be involved in the formation of at least one large-magnitude activity cliff.¹¹ Thus, activity cliffs can be identified in essentially all compound data sets and provide possible starting points for SAR exploration.

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

Received: March 5, 2013

Published: April 14, 2013

MATERIALS AND METHODS

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

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

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

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

Compound data analysis and pathway calculations were carried out with in-house generated Java programs or KNIME¹⁴ protocols.

Pathway Frequency. A normalized pathway frequency was calculated for activity cliff-dependent and -independent

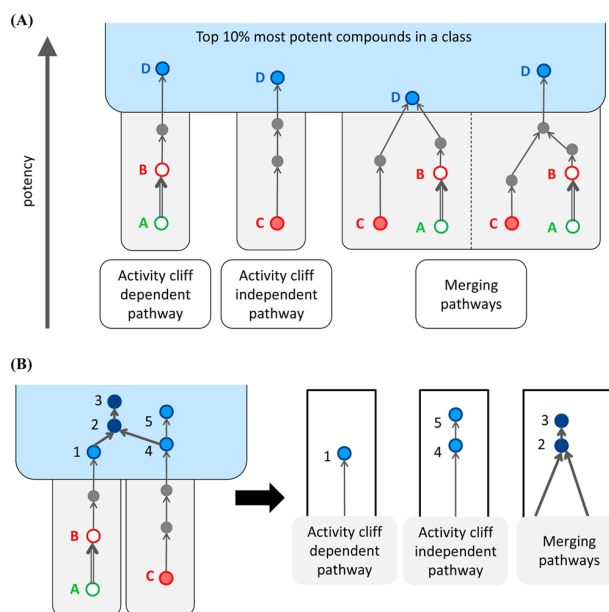


Figure 1. Compound pathways. (A) Schematic illustration of different categories of qualifying potency-directed compound pathways. Highly potent activity cliff partners (B compounds) represent potential starting points of activity cliff-dependent pathways. The weakly potent cliff partners are denoted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

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

Compound Data Sets. To evaluate SAR progression in a systematic manner, we searched ChEMBL¹⁵ (release 14) for target-based compound data sets that evolved over time and in which at least a 10% average frequency for activity cliff-dependent or -independent pathways was observed. Each data set had to contain at least 100 compounds active against a human target with direct interactions (ChEMBL relationship type "D") at the highest confidence level (ChEMBL confidence score "9").¹⁵ In addition, equilibrium constants (K_i values) had to be available as potency measurements for all data set compounds. If several K_i values were available for a compound, the most recent measurement was used. To account for data set evolution over time, the condition was applied that compounds comprising a set had to be reported in increments over a period of at least five subsequent years. During each year, the addition of a new compound subset was required.

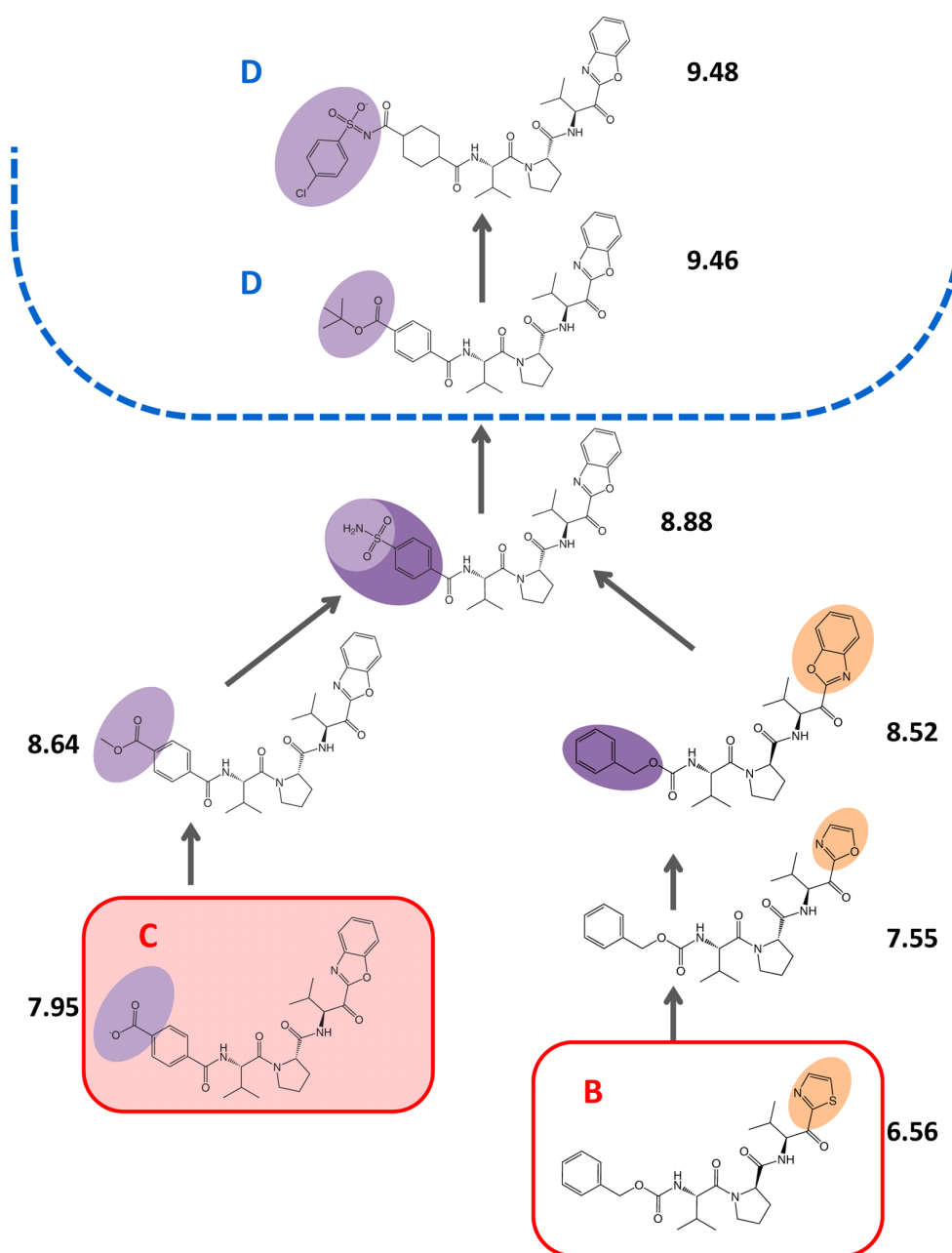


Figure 2. Exemplary pathway. Example of a merging pathway from the leukocyte elastase inhibitor data set shown in detail. Pathway starting points are color-coded according to Figure 1. Structural changes between pairs of compounds in a pathway are highlighted using corresponding colors, and compounds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

RESULTS AND DISCUSSION

Pathway Model of SAR Progression. Our compound pathway model was designed to capture SAR progression in a compound activity class. Series of structurally analogous compounds with steady potency progression toward the most potent compounds found in a data set served as a model for compound optimization paths. Therefore, structural relationships between compounds were established on the basis of sequentially overlapping MMPs with size-restricted structural changes. Furthermore, the formation of pathways was limited to series of increasingly potent compounds that were reported during the same or subsequent years. Hence, a more potent

analog in a pathway was not permitted to be reported earlier than a less potent one. This additional restriction was introduced to model compound optimization over time. Pathways were differentiated according to their origins and potential overlap, as further discussed below.

Activity Cliff-Dependent and -Independent Pathways.

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

Table 1. Evolving Compound Data Sets and Pathway Statistics^a

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

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

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

and -independent pathways because their *D* compounds can be separately reached by both pathway categories. Figure 2 shows an exemplary merging pathway in detail.

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

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

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

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

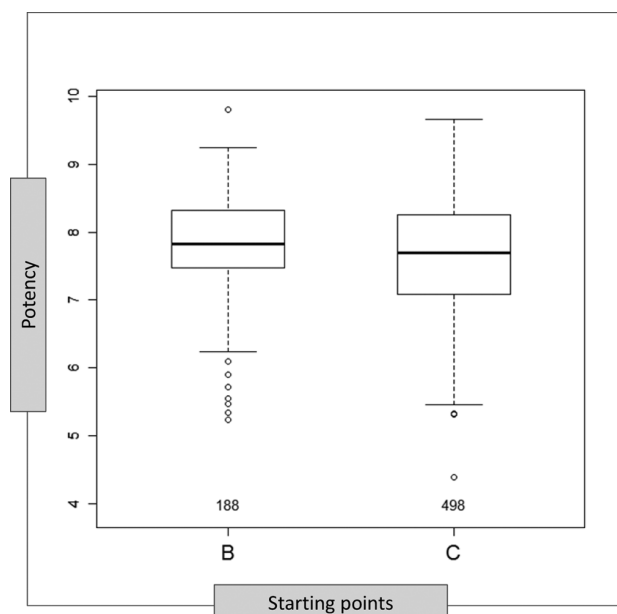


Figure 3. Potency distribution of pathway starting points. Potency distribution of the pathway starting points *B* and *C* from all data sets in boxplots.

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

Pathway Comparison. Given the general differences in the number of *B* and *C* compounds, pathway utilization was quantified and compared by calculating the relative frequency of activity cliff-dependent and -independent pathways with SAR

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

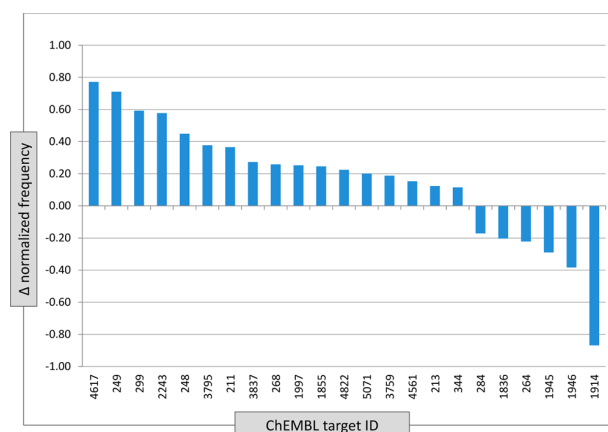


Figure 4. Pathway frequency difference. Difference between normalized frequencies of activity cliff-dependent and -independent pathways targets reported for all 23 data sets with a more than 10% difference. Positive and negative differences indicate larger frequencies of activity cliff-dependent and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

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

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

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

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

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

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

■ ASSOCIATED CONTENT

📄 Supporting Information

Table S1 reports statistics for type B, C, and D compounds and their pathway engagement. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

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

Author Contributions

†The contributions of these authors should be considered equal.

Notes

The authors declare no competing financial interest.

■ REFERENCES

- (1) Peltason, L.; Bajorath, J. Systematic computational analysis of structure-activity relationships: Concepts, challenges and recent advances. *Future Med. Chem.* **2009**, *1*, 451–466.
- (2) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity landscape representations for structure-activity relationship analysis. *J. Med. Chem.* **2010**, *53*, 8209–8223.

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

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

- (5) Vogt, M.; Huang, Y.; Bajorath, J. From activity cliffs to activity ridges: Informative data structures for SAR analysis. *J. Chem. Inf. Model.* **2011**, *51*, 1848–1856.

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

- (7) Medina-Franco, J. L.; Martínez-Mayorga, K.; Bender, A.; Marín, R. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. Characterization of activity landscapes using 2D and 3D similarity methods: Consensus activity cliffs. *J. Chem. Inf. Model.* **2009**, *49*, 477–491.

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

- (9) Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J. MMP-Cliffs: Systematic identification of activity cliffs on the basis of matched molecular pairs. *J. Chem. Inf. Model.* **2012**, *52*, 1138–1145.

- (10) Wassermann, A. M.; Dimova, D.; Bajorath, J. Comprehensive analysis of single- and multi-target activity cliffs formed by currently available bioactive compounds. *Chem. Biol. Drug Des.* **2011**, *78*, 224–228.

- (11) Stumpfe, D.; Bajorath, J. Frequency of occurrence and potency range distribution of activity cliffs in bioactive compounds. *J. Chem. Inf. Model.* **2012**, *52*, 2348–2353.

- (12) Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Cheminformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2004; pp 271–285.

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

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

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

Summary

Herein, we have introduced a compound pathway model to capture positive SAR progression in compound series evolving over time. Pathways originating from activity cliffs (cliff-dependent), starting from any other data set compound (cliff-independent), and combining compounds from activity cliff-dependent and -independent pathways (merging) were systematically determined. Our results revealed that cliff-dependent pathways with positive SAR progression occurred with higher relative frequency than cliff-independent ones. Furthermore, on average, 72% of the most potent data set compounds were reached by pathways. In addition, 53.9% and 28.1% of all cliff-dependent and cliff-independent pathways have reached the most potent compounds, respectively. Taken together, our findings provided clear evidence for improved SAR progression originating from activity cliffs. My contributions to this study have been the implementation and analysis of the pathway model.

Having found an indication for SAR information gain associated with activity cliffs, we have next investigated the question whether or not we can find evidence to what extent activity cliffs are utilized as starting points in practical compound optimization efforts.

Chapter 7

Do Medicinal Chemists Learn from Activity Cliffs? A Systematic Evaluation of Cliff Progression in Evolving Compound Data Sets

Introduction

Activity cliffs are associated with high SAR information content and their study is of major importance for practical medicinal chemistry. Although the concept of activity cliffs is becoming increasingly attractive, it is currently unknown to what extent SAR information provided by activity cliffs is utilized in compound optimization efforts. This is a nontrivial question and difficult to address, especially from a computational perspective. However, providing answers to this question would provide an indication of the proportion of currently unexplored activity cliffs and potentially point of further opportunities for activity cliff analysis in medicinal chemistry. To investigate this question, we have carried out a comprehensive analysis of publicly available compound data sets that evolved

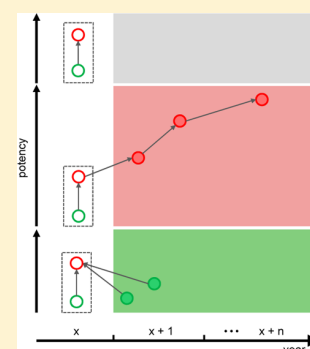
over time and systematically searched for analogues of cliff-forming compounds with increasing potency.

Do Medicinal Chemists Learn from Activity Cliffs? A Systematic Evaluation of Cliff Progression in Evolving Compound Data Sets

Dilyana Dimova,[†] Kathrin Heikamp,[†] Dagmar Stumpfe, and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstrasse 2, D-53113 Bonn, Germany

ABSTRACT: Activity cliffs are defined as pairs of structurally similar compounds with a significant difference in potency. These compound pairs have high SAR information content because they represent small structural changes leading to large potency alterations. Accordingly, activity cliffs are of prime interest for SAR exploration and compound optimization. It is currently unknown to what extent activity cliff information is utilized in practical medicinal chemistry. Therefore, we have assembled 56 compound data sets that evolved over time and searched for analogues of activity cliff-forming compounds with further increased potency. For ~75% of all activity cliffs, there was no evidence for further chemical exploration. For ~25% of all cliffs, potency progression was detected. In total, for ~15% of all activity cliffs, positive cliff progression was observed that often involved multiple analogues. Given these findings, chemically unexplored activity cliffs should provide significant opportunities for further study in medicinal chemistry.



INTRODUCTION

The study of activity cliffs has experienced increasing interest in medicinal chemistry.^{1,2} Activity cliffs have been defined as pairs of structurally similar or analogous compounds with large potency differences.¹ Such compound pairs are an immediate source of SAR information, which makes the activity cliff concept attractive for medicinal chemistry.² In addition, activity cliffs represent the most prominent features of activity landscape representations³ that are modeled to make SARs contained in compound data sets graphically accessible.

Activity cliffs have been systematically surveyed in publicly available compounds, and cliffs were consistently found in sets of compounds with activity against a wide range of targets.^{4,5} In a detailed search for activity cliffs in currently available active compounds, ~5% of all pairs of structural analogues formed activity cliffs spanning a potency difference of at least 2 orders of magnitude.⁵ More than 20% of all active compounds participated in these cliffs, albeit at different frequencies. Hence, although activity cliffs were widely distributed, only a small percentage of all pairs of structurally analogous compounds formed cliffs of significant magnitude. On the other hand, on average, every fifth compound was involved in at least one activity cliff. Thus, activity cliffs were often available in compound data sets as potential starting points for SAR exploration.

There is little doubt that the study of activity cliffs is relevant from a data analysis perspective, taking statistical, information-theoretic, and/or SAR criteria into account. But what about the practice of medicinal chemistry? Activity cliffs are certainly considered in hit-to-lead and lead optimization. Intuitively and on the basis of casual individual experiences, one would assume that activity cliffs are given serious consideration in lead optimization projects. However, is there an objectively measurable general impact on optimization efforts? Can one find firm evidence

for the generation of increasingly potent compounds originating from activity cliffs? Analyzing these questions would help to further evaluate the relevance of the activity cliff concept for medicinal chemistry. Currently, there is little, if any, information available concerning the general impact of activity cliff information on compound optimization.

In order to explore this question in a systematic way, we have assembled compound data sets evolving over time and identified activity cliffs formed by structural analogues. We then determined for all qualifying activity cliffs whether analogues of potent cliff partners became available over time and studied the potency distribution of analogues. For a thorough assessment of activity cliff progression, different categories of cliffs were introduced. The results of our analysis are reported herein.

MATERIALS AND METHODS

Compound Data Sets. Target-based compound data sets evolving over time were selected from ChEMBL,⁶ version 14. Candidates for evolving data sets were prescreened for the presence of clearly defined activity cliffs, as further specified below. Only compounds active against human targets at the highest confidence level (ChEMBL confidence score 9) with reported direct binding interactions (ChEMBL relationship type D) and with equilibrium constants (K_i) available as potency measurements were selected. If several K_i values were reported for a compound, the most recent measurement was used. A qualifying target-based compound set had to contain at least 100 compounds reported in increments over a period of at least 5 subsequent years (i.e., for each year, addition of a subset of new compounds was required). Periods over which evolving compound data sets were monitored ranged from 5 to 18 years. The progression of an exemplary data set over time is reported in Table 1, and all 56 data sets are listed in Table 2.

Received: January 29, 2013

Published: March 25, 2013

Table 1. Evolving Data Set^a

year	compd	Carbonic Anhydrase I Inhibitors (ChEMBL target no. 261)						
		potency range (pK _i)	MMP	activity cliff	hpCP	CAT I	CAT II	CAT III
1999	12	4.7–7.9	6	3	2	1	0	1
2004	107	1.3–9.2	164	17	9	0	2	7
2005	291	0.04–8.7	610	106	34	7	20	7
2006	89	2.6–8.1	231	49	10	2	4	4
2007	77	1.6–8.5	185	35	13	6	6	1
2008	88	1.9–8.8	246	63	14	3	4	7
2009	154	0.9–9.0	655	97	34	16	9	9
2010	194	0.5–7.7	707	41	7	5	1	1
2011	274	1.2–8.5	1432	159	44	44	0	0

^aA representative example of a data set evolving over time is given (carbonic anhydrase I inhibitors). For each of 9 years, the number of newly reported compounds and their potency range are provided as well as the number of MMPs and activity cliffs these compounds formed. In addition, the number of highly potent cliff partners (hpCPs) the activity cliffs contained is given. According to the definition in the text, hpCPs were classified into three categories (CAT I, CAT II, and CAT III). For consistency, all 56 evolving data sets we assembled included all compounds reported during 2011 as the final year.

Size-Restricted Matched Molecular Pairs. A matched molecular pair (MMP)⁷ is defined as a pair of compounds that only differ at a single site, i.e., that can be interconverted by the exchange of a fragment. An exemplary MMP is shown in Figure 1. For all target sets, MMPs were systematically generated using an in-house implementation of the Hussain and Rea algorithm.⁸ For MMP generation, size restrictions were applied to the common core structure of compounds forming an MMP and the exchanged fragments, as reported previously.⁹

The size of exchanged fragments was limited to a maximum of 13 non-hydrogen atoms, and the maximal difference in size between fragments was limited to eight non-hydrogen atoms. In addition, the core structure of MMP-forming compounds was required to have at least twice the size of the larger fragment. These upper-limit size restrictions were introduced to ensure that chemical replacements were small and chemically intuitive. Specifically, given these restrictions, the largest permitted substitutions would include the addition of a substituted six-membered ring to a compound or the replacement of a 5- or 6-membered ring by a substituted 2-ring system containing a maximum of 10 ring atoms.

If alternative substructure exchanges were possible to generate an MMP, the smallest structural transformation was applied. Introducing these size limitations ensured that MMPs consisted of closely related structural analogues.⁹ This protocol was applied to all 56 target sets and resulted in the generation of 370–15890 MMPs per set.

Activity Cliff Criteria. For a consistent description of activity cliffs, two criteria must be clearly defined including compound similarity and the potency difference threshold. We followed the definition of MMP-cliffs;⁹ i.e., two compounds qualified as structurally similar if they formed a size-restricted MMP. This structurally conservative definition of activity cliffs avoided ambiguities associated with calculation of numerical similarity values on the basis of molecular descriptors.² As a potency difference threshold, a difference in equilibrium constants of at least 2 orders of magnitude was required, which limited the analysis to activity cliffs of significant magnitude⁵ and avoided potential bias due to the use of approximate activity measurements.⁵ An exemplary MMP-cliff is shown in Figure 1. The number of activity cliffs per compound set ranged from 51 to 1236.

Activity Cliff Categorization. Activity cliffs were systematically identified in all evolving data sets, and the highly potent cliff partners (hpCPs) were subjected to chemical neighborhood analysis over time by searching for structural analogues. The search was carried out on the basis of size-restricted MMPs; i.e., all subsequently reported compounds were identified that formed MMPs with a highly potent cliff partner. Then each activity cliff was assigned to one of three categories (CAT), as illustrated in Figure 2A.

- (1) CAT I: No structural analogues of the hpCP were found in evolving compound data sets.
- (2) CAT II: One or more structural analogues of the hpCP with further increased potency were identified.

- (3) CAT III: One or more structural analogues of the hpCP were found that did not display potency improvements or had reduced potency.

Compound data analysis and activity cliffs calculations were carried out with in-house generated Java programs or KNIME¹⁰ protocols.

RESULTS AND DISCUSSION

Analysis Concept. How can one obtain evidence for the potential use of activity cliff information in medicinal chemistry? This is a nontrivial question, especially if one would like to explore it systematically. It is not possible to quantify medicinal chemists' intuition of how to best approach a compound optimization task. Neither is it possible to determine with certainty whether activity cliffs have been taken into consideration nor whether the notion of activity cliff(s) has served as a starting point for an optimization effort.

However, it is possible to determine for each activity cliff detected in a data set whether or not structural analogues of potent cliff-forming compounds have subsequently been reported. If SAR information provided by an activity cliff has been utilized, analogues of the potent cliff partner might become available. If this is not the case, i.e., if an activity cliff has remained "isolated", it can be concluded with certainty that the cliff has not been further chemically explored. If analogues are available, the situation is different. Although it cannot be confirmed that cliff compounds have provided an immediate starting point for the design of these analogues, it can be concluded with certainty that the structural neighborhood of a cliff has been explored, a process we refer to as "activity cliff progression". Although lead optimization is a multiparametric process, the primary goal of exploring activity cliff information is further increasing compound potency, consistent with the SAR information provided by a cliff. Accordingly, one can distinguish two principal outcomes of activity cliff progression, i.e., either analogues of potent cliff partners display further increased potency or not. The former case represents the desirable positive result for consideration of activity cliffs in the context of SAR analysis; i.e., the structural neighborhood of an activity cliff has been explored, leading to the generation of compounds with further increased potency.

On the basis of the above considerations, activity cliffs have been categorized for the analysis of cliff progression, as detailed in the section Materials and Methods. CAT I represents isolated

Table 2. Data Sets and Category Distribution^a

no.	ChEMBL target no.	target	compd	hpCP	CAT I		CAT II		CAT III	
					hpCP	%	hpCP	%	hpCP	%
1	244	coagulation factor X	1198	313	225	71.9	82	26.2	6	1.9
2	237	κ opioid receptor	1304	185	94	50.8	64	34.6	27	14.6
3	261	carbonic anhydrase I	1286	167	84	50.3	46	27.5	37	22
4	256	adenosine A3 receptor	1710	178	84	47.2	46	25.8	48	27
5	205	carbonic anhydrase II	1361	118	56	47.5	33	28.0	29	24.6
6	3594	carbonic anhydrase IX	946	85	48	56.5	28	32.9	9	10.6
7	249	neurokinin 1 receptor	211	47	21	44.7	24	51.1	2	4.3
8	233	μ opioid receptor	1391	153	111	72.5	21	13.7	21	13.7
9	259	melanocortin receptor 4	1273	147	113	76.9	20	13.6	14	9.5
10	4617	phenylethanolamine N-methyltransferase	148	35	6	17.1	19	54.3	10	28.6
11	264	histamine H3 receptor	1515	146	98	67.1	18	12	30	20.5
12	253	cannabinoid CB2 receptor	1403	180	143	79.4	18	10	19	10.6
13	234	dopamine D3 receptor	881	115	88	76.5	15	13	12	10.4
14	1855	gonadotropin-releasing hormone receptor	267	59	35	59.3	14	23.7	10	16.9
15	251	adenosine A2a receptor	2114	208	179	86.1	13	6.3	16	7.7
16	1914	butyrylcholinesterase	159	29	11	37.9	13	44.8	5	17.2
17	238	dopamine transporter	628	52	38	73.1	11	21.2	3	5.8
18	3759	histamine H4 receptor	334	47	36	76.6	11	23.4	0	0.0
19	228	serotonin transporter	1099	59	43	72.9	10	16.9	6	10.2
20	218	cannabinoid CB1 receptor	1320	85	63	74.1	9	10.6	13	15.3
21	204	thrombin	808	93	82	88.2	8	8.6	3	3.2
22	344	melanin-concentrating hormone receptor 1	870	124	98	79.0	8	6.5	18	14.5
23	217	dopamine D2 receptor	1431	70	52	74.3	8	11.4	10	14.3
24	222	norepinephrine transporter	853	36	19	52.8	8	22.2	9	25
25	4822	β -secretase 1	116	22	12	54.5	8	36.4	2	9.1
26	226	adenosine A1 receptor	1859	120	102	85.0	7	5.8	11	9.2
27	4308	bradykinin B1 receptor	415	114	101	88.6	5	4.4	8	7.0
28	1800	corticotropin releasing factor receptor 1	477	85	78	91.8	5	5.9	2	2.4
29	214	serotonin 1a (5-HT1a) receptor	938	63	53	84.1	5	7.9	5	7.9
30	255	adenosine A2b receptor	798	57	46	80.7	4	7	7	12.3
31	284	dipeptidyl peptidase IV	276	26	19	73.1	4	15.4	3	11.5
32	268	cathepsin K	272	48	44	91.7	3	6.3	1	2.1
33	2014	nociceptin receptor	599	115	110	95.7	3	2.6	2	1.7
34	236	δ opioid receptor	1197	101	92	91.1	3	3	6	5.9
35	2954	cathepsin S	371	85	81	95.3	3	3.5	1	1.2
36	3798	calcitonin gene-related peptide type 1 receptor	246	29	23	79.3	3	10.3	3	10.3
37	1836	prostanoid EP4 receptor	194	28	25	89.3	3	10.7	0	0.0
38	3371	serotonin 6 (5-HT6) receptor	888	34	27	79.4	2	5.9	5	14.7
39	1997	equilibrative nucleoside transporter 1	118	13	11	84.6	2	15.4	0	0
40	245	muscarinic acetylcholine receptor M3	343	53	51	96.2	1	1.9	1	1.9
41	1889	vasopressin V1a receptor	318	43	42	97.7	1	2.3	0	0
42	213	β -1 adrenergic receptor	175	29	28	96.6	1	3.4	0	0
43	219	dopamine D4 receptor	436	22	17	77.3	1	4.5	4	18.2
44	5071	G-protein coupled receptor 44	376	54	54	100	0	0	0	0
45	4561	neuropeptide Y receptor type 5	234	51	51	100	0	0	0	0
46	299	protein kinase C alpha	168	35	33	94.3	0	0	2	5.7
47	224	serotonin 2a (5-HT2a) receptor	657	34	34	100	0	0	0	0
48	3837	cathepsin L	201	34	34	100	0	0	0	0
49	1946	melatonin receptor 1B	262	28	27	96.4	0	0	1	3.6
50	2243	anandamide amidohydrolase	101	28	28	100	0	0	0	0
51	211	muscarinic acetylcholine receptor M2	199	24	22	91.7	0	0	2	8.3
52	231	histamine H1 receptor	187	24	24	100	0	0	0	0
53	240	HERG	619	23	22	95.7	0	0	1	4.3
54	248	leukocyte elastase	192	18	18	100	0	0	0	0
55	3795	melanocortin receptor 1	134	17	15	88.2	0	0	2	11.8
56	1945	melatonin receptor 1A	215	15	15	100	0	0	0	0

^aAll 56 evolving compound data sets are listed in the order of decreasing numbers of CAT II hpCPs. Target names and ChEMBL target identifications are provided. Abbreviations are used according to Table 1. For each class, the distribution of hpCPs over CATs I, II, and III is reported.

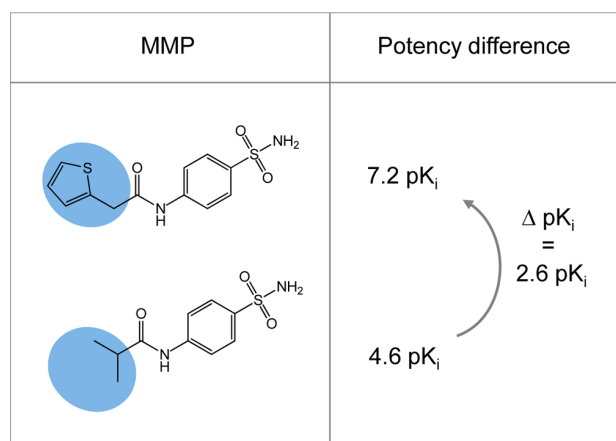


Figure 1. MMP-cliff. Two small inhibitors of carbonic anhydrase 1 are shown that form a size-restricted MMP. The distinguishing fragments are highlighted. The potency difference between these compounds is larger than 2 orders of magnitude, and hence, the two compounds form an activity cliff (MMP-cliff). The compound at the top is the highly potent cliff partner (hpCP).

cliffs, whereas CAT II and CAT III represent different cases of cliff progression.

In addition, another point should be considered. Activity cliffs can be defined in a number of different ways including the application of alternative structural/similarity and potency difference criteria.² However, because our cliff progression analysis is based on the detection of compound analogue series, the application of a well-defined structure-based definition of activity cliffs such as MMP-cliffs is essential.

Activity Cliff Progression. In the exemplary data set reported in Table 1, a steady increase in the number of compounds, MMPs, and activity cliffs occurred over time. During each recorded year, newly formed activity cliffs were identified. In compound subsets that became available during the following years, an MMP-based search for analogues of each highly potent cliff partner was carried out to categorize these cliffs. For example, from 2005 to 2006, there was an increase in the number of unique activity cliffs from 126 to 175. The 49 new cliffs that became available during 2006 involved 10 new hpCPs. For these potent cliff partners, analogue searches were carried out using all subsequently reported compounds. The generally observed difference in growth rates between activity cliffs and hpCPs was due to the fact that an hpCP typically formed multiple activity cliffs with weakly potent analogues (Table 1). Since our analysis of activity cliff progression focused on potency changes relative to each hpCP (rather than weakly potent cliff partners), all cliffs formed by an hpCP had to be considered only once.

In the evolving carbonic anhydrase I inhibitor set in Table 1, a total of 167 hpCPs were available at the 2011 end point, which formed 84 CAT I, 46 CAT II, and 37 CAT III cliffs. Hence, about half of these activity cliffs were isolated and no structural evidence was found for further exploration. By contrast, analogues were identified for 83 other hpCPs and in 46 of these cases, analogues had further increased potency. Figure 2B shows representative examples of CAT II cliff progression.

In our analysis of all data sets, as discussed in the following, we identified only 10 hpCPs (CAT III) for which all available analogues had conserved potency. Given this very small number, we considered CAT III and CAT II to represent “negative” and “positive” cliff progression, respectively.

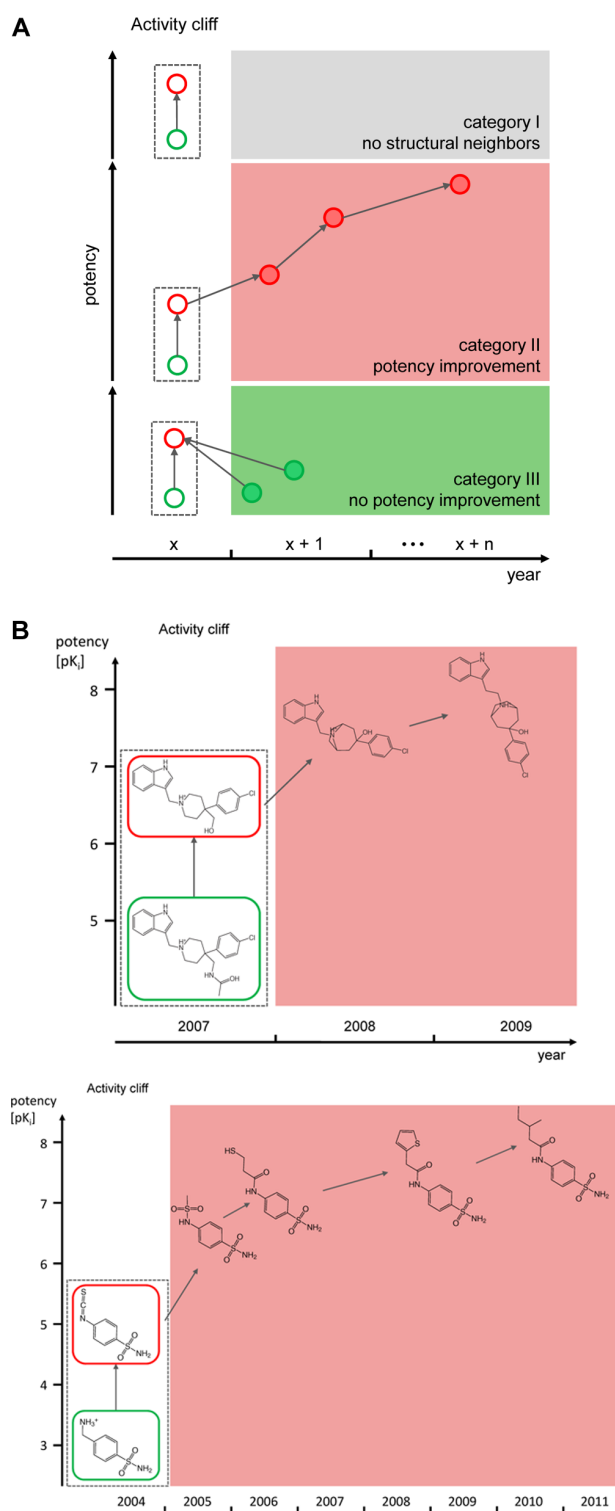


Figure 2. Activity cliff categories. (A) Schematic representation of three different categories of activity cliffs. Compounds are depicted as colored nodes (red, high potency; green, low potency). Activity cliff-forming compounds are represented as unfilled nodes and structural analogues of the highly potent cliff partner (hpCP, red) as filled nodes. The hpCP served as a reference point for activity cliff progression monitored over time. (B) Two representative examples of CAT II cliffs are shown including dopamine D3 receptor antagonists and carbonic anhydrase I inhibitors.

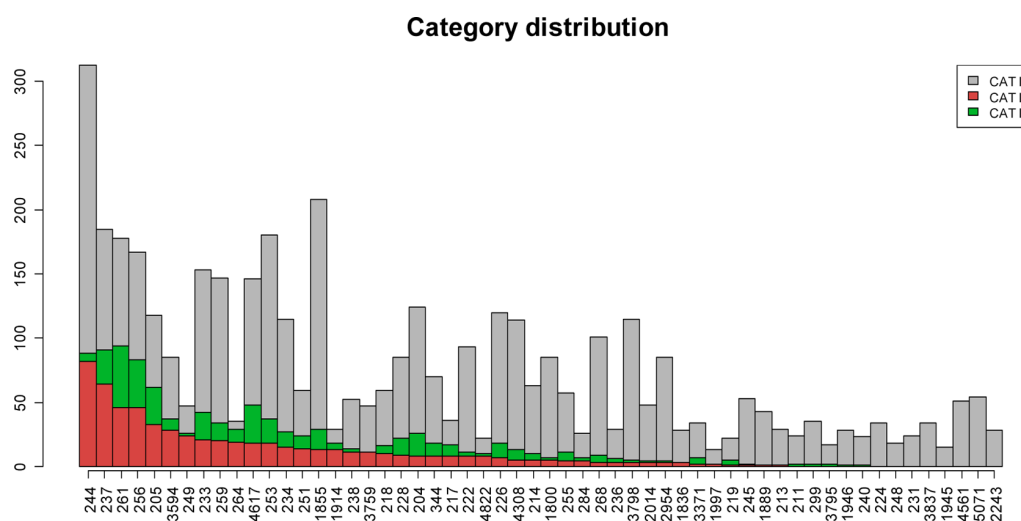


Figure 3. Category distribution. For each compound data set, the absolute numbers of hpCPs are reported that form CAT I (gray), CAT II (red), or CAT III (green) activity cliffs. Bars represent cumulative counts (i.e., red + green + gray). For data sets, ChEMBL target identifications are provided. For example, for the first data set (ChEMBL target no. 244), there are 82 CAT II, six CAT III, and 225 CAT I cliffs.

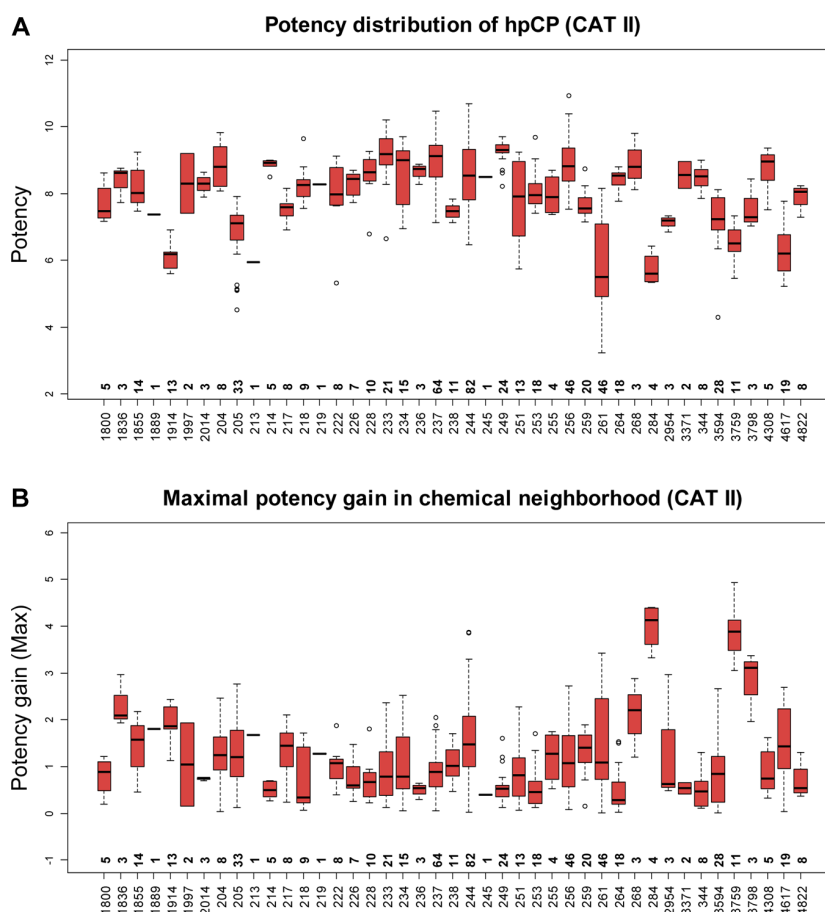


Figure 4. Potency distribution and maximal gain. For all target sets, box plots are shown that monitor the (A) potency distribution of hpCPs from CAT II cliffs and (B) maximal potency gain during activity cliff progression. For data sets, ChEMBL target identifications are provided. At the bottom of each box plot, the number of CAT II hpCPs per data set is reported.

Large-Scale Progression Analysis. In Table 2, cumulative results are reported for the 2011 end point of all 56 evolving compound sets. We note that hpCPs were frequently occurring

in these data sets, ranging from 13 for equilibrative nucleoside transporter 1 ligands to 313 for coagulation factor X inhibitors, which provided a sound basis for cliff categorization and

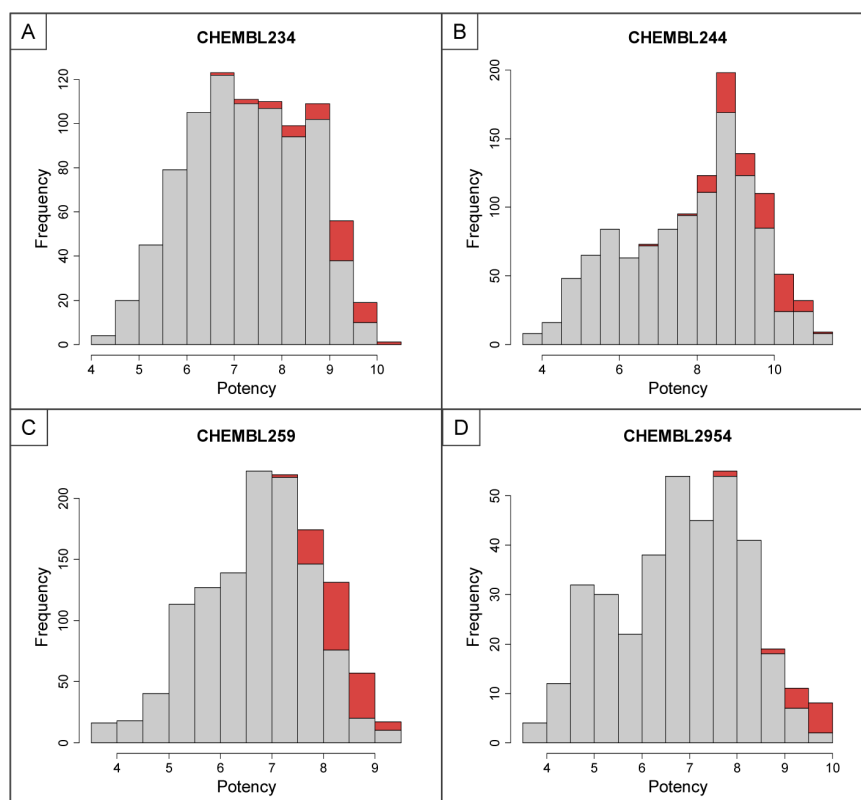


Figure 5. Compound frequencies and potency ranges. The number and the potency range distribution of compounds not involved in positive activity cliff progression (gray) and of all compounds associated with CAT II cliffs (red) are compared. Histograms are shown for four exemplary compound data sets: (A) dopamine D3 receptor antagonists (ChEMBL target no. 234), (B) coagulation factor X inhibitors (no. 244), (C) melanocortin receptor 4 antagonists (no. 259), and (D) cathepsin S inhibitors (no. 2954).

progression analysis across many different targets. In Figure 3, the distribution of cliff categories is monitored over all data sets to complement the statistics provided in Table 2. Several observations are made. In the majority of compound data sets (53 of 56), isolated activity cliffs were more frequent than progressing cliffs, regardless of absolute cliff numbers in the sets. In eight data sets, only CAT I cliffs were observed. Thus, as illustrated in Figure 3, for the majority of all hpCP activity cliffs (~75%), no evidence for further chemical exploration was detectable. By contrast, in 48 compound sets, at least limited activity cliff progression was observed. In these cases, CAT II cliffs (positive progression) were overall more frequent (total count 611) than CAT III cliffs (negative progression, total count 426). In 43 of 56 sets, positive cliff progression was detected. The examples of CAT II cliff progression shown in Figure 2B were representative of many cases. Single more potent analogues of an hpCP were frequently detected as well as series of analogues with gradually increasing potency indicative of optimization paths. In total, for ~15% of all activity cliffs, hpCP analogues with further increased potency were identified.

Potency Distribution. Despite SAR information associated with activity cliffs, it might often be difficult to translate this information into compounds with further increased potency if activity cliffs already involve highly potent compounds. This point should also be taken into account in the analysis of activity cliff progression. Although we determined activity cliffs over the entire potency range, many hpCPs of progressing cliffs were potent in the mid to low nanomolar range. In Figure 4A, the potency distribution of CAT II hpCPs is monitored over all 43

compound sets containing CAT II cliffs. In 25 cases, the median potency of hpCPs was 10 nM or higher. In Figure 4B, the distribution of the maximal potency gain for CAT II cliffs is reported for all 43 data sets. In 21 and 19 cases, median potency increases of up to 1 order of magnitude and of 1–2 orders of magnitude were observed, respectively. In three instances, potency increases of 3–4 orders of magnitude were detected. In addition, individual analogues with large increases in potency relative to their hpCPs were often found. Given the prevalence of highly potent hpCPs, as revealed in Figure 4A, the frequently observed 10- to 100-fold increases in compound potency during CAT II cliff progression were considered significant. In Figure 5, the potency distribution of all compounds involved in positive activity cliff progression versus all others is illustrated for four exemplary data sets. In each case, compounds associated with CAT II cliffs were among the most potent compounds in the data set.

Concluding Remarks. We have investigated the question to what extent activity cliff information might be utilized in medicinal chemistry, which is not straightforward to address. The approach taken was based on the concept of activity cliff progression. Specifically, if activity cliffs reveal SAR determinants and if such insights are considered in medicinal chemistry, then one should expect to observe a further exploration of the immediate chemical neighborhood of cliffs. In successful cases, this exploration should result in analogues of activity cliff compounds with further increased potency. To investigate activity cliff progression in a systematic manner, we assembled data sets evolving over time, identified all activity cliffs, and searched for subsequently reported analogues of potent cliff partners. For systematic analysis of activity

cliff progression, a categorization of cliffs was introduced. Care was taken to limit the analysis to clearly defined activity cliffs. Therefore, only activity cliffs were considered that consisted of structural analogues with a potency difference of at least 2 orders of magnitude on the basis of equilibrium constants as activity measurements. In our analysis, we found no structural analogues of potent cliff partners for ~75% of the activity cliffs across all compound sets. Hence, there was no evidence that these cliffs were chemically further explored. On the basis of these findings, one would conclude that activity cliff information is currently underutilized in the practice of medicinal chemistry, despite the substantial interest the activity cliff concept has been experiencing. There are likely reasons for this. Activity cliffs do not represent an a priori and immediately accessible data structure for medicinal chemistry. Rather, they are subject to clear definition, systematic exploration, and importantly, representation in a chemically intuitive format. Challenges associated with these requirements might currently limit the utility of the activity cliff concept for lead optimization. On the other hand, for ~25% of all activity cliffs, progression was detected and for ~15% of all cliffs, analogues of potent cliff partners with potency increases of often 1 or 2 orders of magnitude were identified. In total, positive progression was confirmed for more than 600 activity cliffs distributed over 43 different target sets. For a subset of these cliffs, multiple analogues with gradually increasing potency were observed, forming apparent optimization paths.

Taken together, the results indicate that activity cliffs currently are not a major focal point of practical medicinal chemistry efforts. On average, there is evidence for the exploration of the chemical neighborhood of only every fourth activity cliff present in diverse data sets having evolved over time. However, positive activity cliff progression is observed for every sixth to seventh cliff, and in these cases, structural analogues of activity cliff compounds with significantly increased potency are available. In light of these findings, activity cliffs with as of yet unexplored chemical neighborhoods should provide many opportunities for further exploitation of activity cliff information focusing on compound potency improvement.

AUTHOR INFORMATION

Corresponding Author

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

Author Contributions

†The contributions of these authors should be considered equal.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

D.S. is supported by Sonderforschungsbereich 704 of the Deutsche Forschungsgemeinschaft.

ABBREVIATIONS USED

CAT, category; hpCP, highly potent (activity) cliff partner; MMP, matched molecular pair; SAR, structure–activity relationship

REFERENCES

- (1) Maggiora, G. M. On Outliers and Activity Cliffs—Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535–1535.
- (2) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J. Med. Chem.* **2012**, *55*, 2932–2942.

- (3) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity Landscape Representations for Structure–Activity Relationship Analysis. *J. Med. Chem.* **2010**, *53*, 8209–8223.

- (4) Wassermann, A. M.; Dimova, D.; Bajorath, J. Comprehensive Analysis of Single- and Multi-Target Activity Cliffs Formed by Currently Available Bioactive Compounds. *Chem. Biol. Drug Des.* **2011**, *78*, 224–228.

- (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) 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.

- (7) Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2004; pp 271–285.

- (8) 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.

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

- (10) Tiwaria, A.; Sekhar, A. K. T. Workflow Based Framework for Life Science Informatics. *Comput. Biol. Chem.* **2007**, *31*, 305–319.

Summary

In this study, we have investigated the question of whether SAR information encoded in activity cliffs is utilized in medicinal chemistry. The concept of activity cliff progression was introduced to mimic the exploration of the chemical neighborhood of cliff-forming compounds. Activity cliffs were categorized based on whether or not activity cliff progression was detected and whether such progression ultimately led to a potency increase. For nearly 75% of the activity cliffs, there was no evidence for further chemical exploration. Accordingly, for ~25% of the cliffs, activity cliff progression was detected, and in ~15% of all instances the progression introduced a significant potency increase of 1 or 2 orders of magnitude. The findings indicated that activity cliffs were not major focal points in chemical optimization efforts. Hence, the significant number of still unexplored activity cliff neighborhoods should provide many opportunities for further analysis and compound potency improvement. My contributions to this work have been to the design and implementation of the activity cliff progression model.

In the previous studies, key aspects of the activity cliff concept (e.g., their global distribution across different targets, SAR advantage and utility in medicinal chemistry) have been addressed. Another important question related to activity cliffs is their coordination, i.e., whether or not cliff formation involves multiple active compounds and cliffs. A recent systematic survey² has reported that the majority of activity cliffs are coordinated, rather than isolated, yet their composition and topologies have thus far not been investigated. This question has been explored in the next study.

Chapter 8

Composition and Topology of Activity Cliff Clusters Formed by Bioactive Compounds

Introduction

Conventionally, activity cliffs are explored on the basis of individual compound pairs. However, cliffs often occur in a coordinated (i.e., cliff-forming compounds participate in additional cliffs), rather than isolated manner (i.e., two cliff partners form a single cliff).² Higher-order configurations are thought to be more SAR informative than isolated cliffs, however, their compositions and topologies are currently unknown. The study of coordinated cliffs, their size and topological organization further refines the activity cliff concept and shed light on another as of yet unexplored activity cliff facet. To these ends, we have carried out a systematic survey of all activity cliff configurations formed by currently available bioactive compounds. These configurations have been further classified and their frequency of occurrence and target distribution has been determined.

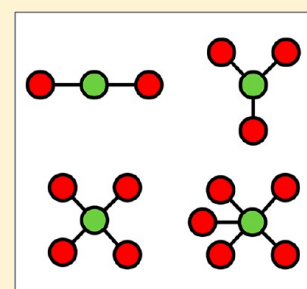
Composition and Topology of Activity Cliff Clusters Formed by Bioactive Compounds

Dagmar Stumpfe, Dilyana Dimova, and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113 Bonn, Germany

Supporting Information

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



INTRODUCTION

The activity cliff concept^{1–4} has experienced increasing attention in computational and medicinal chemistry.^{2–4} Originally, activity cliffs were defined as pairs of structurally similar active compounds having a large difference in potency.^{1,2} Given this definition, the specification of similarity and potency difference criteria is of critical relevance for the assessment of activity cliffs.^{2–4} The popularity of the activity cliff concept in medicinal chemistry is primarily due to the underlying “small chemical changes—large biological effects” paradigm, which assigns high structure–activity relationship (SAR) information content to activity cliffs.^{2,3} In addition to SAR exploration, activity cliffs are of interest for computational analysis because they can be explored through systematic mining of compound activity data^{3,4} and because they are focal points of activity landscape modeling.^{5,6} A variety of molecular representations have been utilized to assess compound similarity in the analysis of activity cliffs, typically in combination with Tanimoto similarity calculations.^{2,3,7} However, in medicinal chemistry, activity cliffs defined on the basis of such whole-molecule similarity calculations are often difficult to interpret.³ Therefore, activity cliffs have also been defined on the basis of substructure relationships between active compounds,³ for example, by employing the matched molecular pair (MMP) formalism.⁸ An MMP is generally defined as a pair of compounds that only differ by a structural change at a single site,^{8,9} i.e., the exchange of two substructures, a so-called chemical transformation.⁹ By introduc-

ing transformation size restrictions,¹⁰ such structural changes can be limited to small and chemically meaningful replacements that relate analogous compounds to each other. The formation of such transformation size-restricted MMPs has been applied as a similarity criterion for activity cliffs, leading to the introduction of MMP-cliffs.¹⁰ The potency difference criterion is also critical for activity cliff analysis. Given the high relevance of activity cliffs for SAR analysis, the exclusive consideration of high-confidence activity data is strongly recommended.^{3,4} A generally preferred activity cliff definition has been put forward that requires the formation of a transformation size-restricted MMP for cliff partners and the presence of a potency difference of at least 2 orders of magnitude on the basis of equilibrium constants (K_i values) as activity measurements.⁴ Activity cliffs have been systematically identified in publicly available compounds active against current targets.^{11,12} Depending on chosen molecular representations, ~20–35% of all compounds with available high-confidence activity data have been found to participate in the formation of at least one well-defined activity cliff, with MMP-cliffs being the structurally most conservative representation of cliffs.¹²

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

Received: December 9, 2013

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

MATERIALS AND METHODS

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

MMP-Cliffs. Activity cliffs were defined as MMP-cliffs¹⁰ with a potency difference of at least 2 orders of magnitude between cliff-forming compounds.⁴ Transformation size-restricted MMPs¹⁰ were systematically generated for all qualifying compounds using an in-house implementation of the Hussain and Rea algorithm.⁹

Network Analysis. All MMP-cliffs were pooled, and a target-based activity cliff network was generated. In this network, nodes represented cliff-forming compounds and edges, activity cliffs. Network representations were drawn with Cytoscape,¹⁵ and network characteristics¹⁶ were assessed. In addition to global network topology and average node degrees, network “heterogeneity” and “centralization” were calculated as parameters related to the neighborhood of a given node.¹⁷ The network heterogeneity is an index accounting for the variance of connectivity and reflecting the tendency of a network to contain hubs.¹⁷ In addition, the network centralization index is a measure of the centrality of nodes; i.e., it describes the extent to which subsets of nodes are more central than others in the network based on their connectivity. For example, the centralization score of networks with a star-like topology is usually close to 1, whereas the score of uniformly connected networks is close to 0.¹⁷ For activity cliff networks, these indices have been calculated using the Cytoscape NetworkAnalyzer plug-in.¹⁸

RESULTS AND DISCUSSION

Activity Cliff Statistics. For activity cliff analysis, all bioactive compounds were selected for which high-confidence activity data were available. No data set size restrictions were applied to ensure maximal target coverage. For the 77 415 qualifying compounds, a total of 20 080 MMP-cliffs were identified in 293 target sets. Many small or very small sets did not yield MMPs or MMP-cliffs. These 20 080 activity cliffs included a total of 18 567 unique cliffs detected in one or more target sets. The number of cliffs exclusively identified in a single target set (intra-class cliffs) was 17 287, whereas only 1280 cliffs were found in more than one set (interclass cliffs). Hence, only 6.9% of all MMP-cliffs are multitarget cliffs. The activity cliff statistics are reported in Table 1. The large number of activity cliffs were formed by 11 783

Table 1. MMP-Cliff Distribution^a

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

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

unique compounds, which yielded 14 044 activity cliff compounds. A total of 1766 compounds with multitarget activity participated in the formation of activity cliffs in different target sets (thus rationalizing the difference between the number of unique and cliff-forming compounds). The number of MMP-cliffs per target set varied between one and 1241. Among the cliff-forming compounds, 7358 exclusively acted as highly potent cliff partners, 6414 exclusively as lowly potent partners, and only 272 compounds were found to participate as highly and lowly potent partners in different activity cliffs. All activity cliffs were subjected to network analysis to visualize the configurations they formed.

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

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

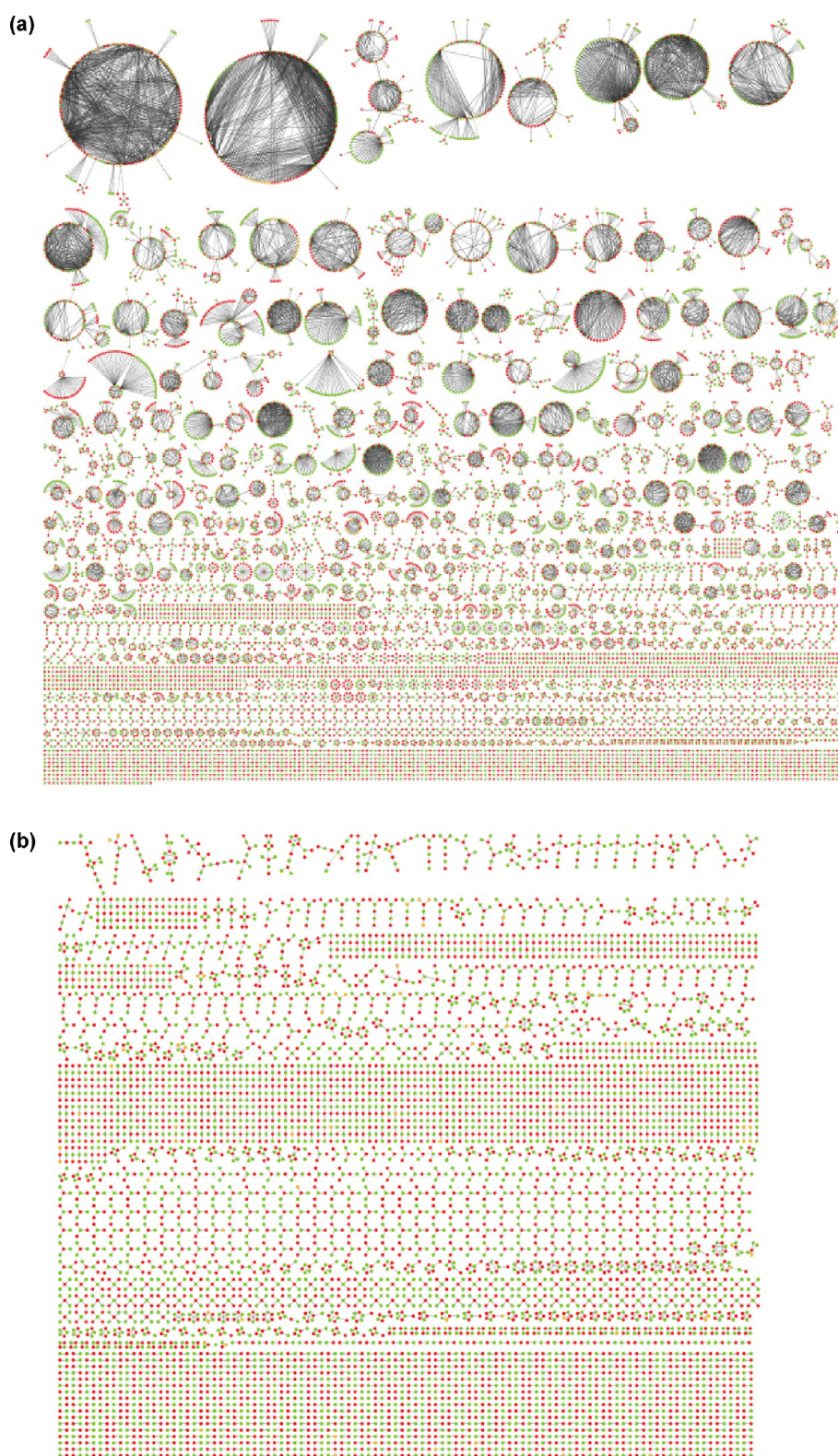


Figure 1. continued

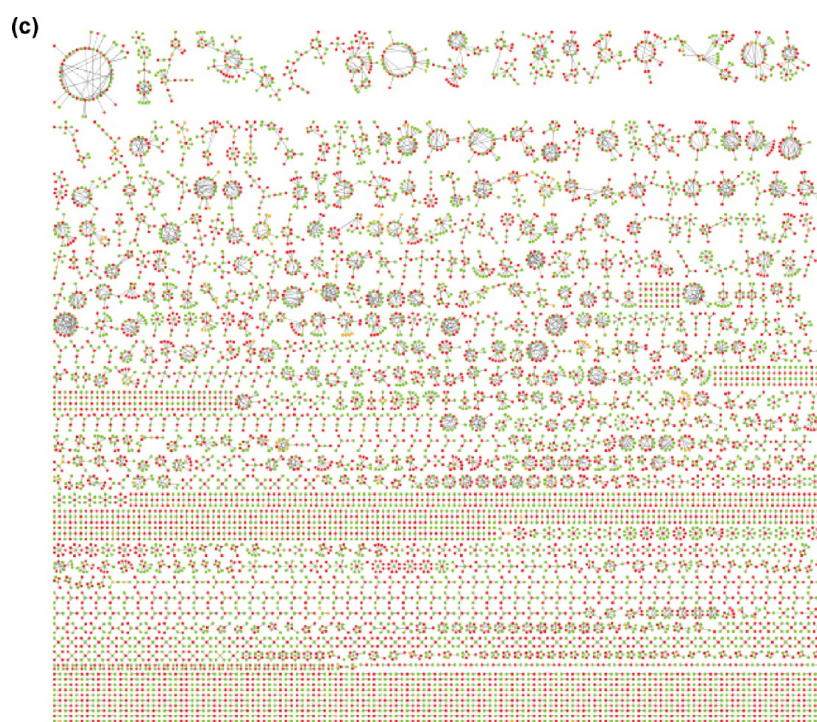


Figure 1. Activity cliff networks. In a, the complete MMP-cliff network is shown. Nodes are colored green if a compound is a highly potent cliff partner, red (lowly potent cliff partner), or yellow if a compound is involved in different activity cliffs both as a highly and lowly potent partner. In b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

Characterization. The activity cliff network contained a total of 2072 differently sized clusters (including isolated cliffs). The cluster order distribution is reported in Table 2. Cluster order

Table 2. Activity Cliff Cluster Order Distribution^a

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

^aThe distribution of activity cliff cluster orders (i.e., numbers of nodes per cluster) across the network is reported.

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

On the basis of global network analysis, we determined that the union of all clusters followed the power law $P(k) \sim k^{-\gamma}$, with γ having a value of 2.5, which is characteristic of *scale-free* networks that typically yield γ values of 2–3.¹⁶ Here, $P(k)$ is the subset of nodes in the network having k connections to others.

Table 3 reports the node degree distribution in the network. Node degrees varied between 1 and 67, with an average node

Table 3. Node Degree Distribution^a

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

^aThe node degree distribution of the activity cliff network is reported.

degree of 2.9. Overall, 1552 nodes with a degree of 5–9 were detected and 496 nodes with a degree of 10–20, revealing the presence of many densely connected nodes. Thus, nodes with a degree ≥ 5 were considered *activity cliff hubs*. In total, the network contained 2166 (15.4%) and 614 (4.4%) hubs with a degree ≥ 5 and a degree ≥ 10 , respectively.

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

network variants were generated after removal of hubs with a degree ≥ 5 and ≥ 10 , respectively. These network variants are displayed in Figure 1b and c. In addition, a comparison of the statistics for the original network and its two variants is presented in Table 4. The absence of increasing numbers of hubs led to

Table 4. Activity Cliff Network Statistics^a

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

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

increasing randomness of the network variants. As expected, these modifications reduced network heterogeneity, a measure for the tendency of a network to contain hubs, and also network centralization, thus indicating increasingly uniform connectivity. Taken together, these findings were also consistent with the global scale-free character of the original activity cliff network.

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

Distribution. Table 5 reports the cluster topology distribution of the activity cliff network. The global network consisted of 2072

Table 5. Cluster Topology Distribution^a

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

^aAll activity cliff clusters are organized according to their topologies and compound composition.

activity cliff clusters that represented 450 distinct cluster topologies. A small set of 39 cluster topologies (including isolated cliffs) accounted for 1630 different clusters. These clusters were of small to moderate size (with up to 12 compounds) and contained a total of 5323 compounds forming 3866 activity cliffs. In addition, 416 clusters with fewer than 50 compounds yielded 385 distinct topologies. Hence, these topologies were detected only once or twice. The corresponding clusters contained 6722 compounds that formed $\sim 55\%$ (11 083) of all activity cliffs. Furthermore, there were 26 large clusters with unique topologies that contained a total of 1999 compounds

forming 5131 cliffs. As further discussed below, large clusters were typically characterized by a high degree of activity cliff density.

The cluster frequency distribution of unique topologies is reported in Table 6. A total of 381 clusters with unique topology

Table 6. Cluster Frequency for Unique Topologies^a

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

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

were observed only once, whereas six and seven distinct topologies were each observed 10 or more and 20 or more times, respectively (i.e., in the latter case, 20 or more clusters were found to share one of seven unique topologies). Thirteen of the 39 topologies that were observed at least three times contained a hub with a degree ≥ 5 .

Topological Categories. We determined that the 1630 clusters with recurrent topology could be assigned to only three main topology categories and a limited number of extensions of these categories, as schematically shown in Figure 2. The main categories included the *star*, *chain*, and *rectangle*

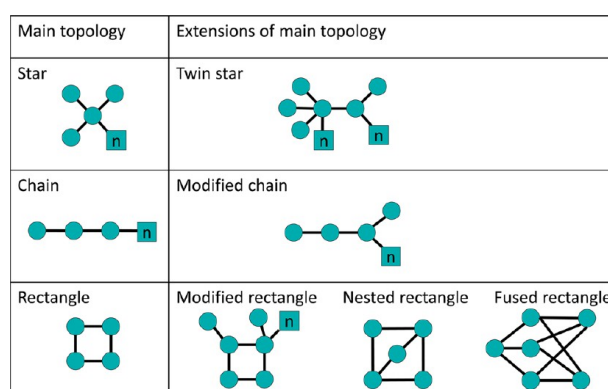


Figure 2. Topology categories. The three most frequently observed activity cliff cluster topologies (left) and their extensions (right) are schematically illustrated. The three main topology categories were termed *star*, *chain*, and *rectangle*, respectively. Squared nodes represent variable node numbers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

topologies. For the *star* and *chain* topologies, frequently observed extensions included the *twin star* and *modified chain*, respectively. In addition, the *rectangle* topology had three well-defined extensions including the *modified*, *nested*, and *fused rectangle*, as illustrated in Figure 2. Taken together, this limited set of topologies or combinations of these topologies covered all small and moderately sized activity cliff clusters for compounds active against current targets.

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

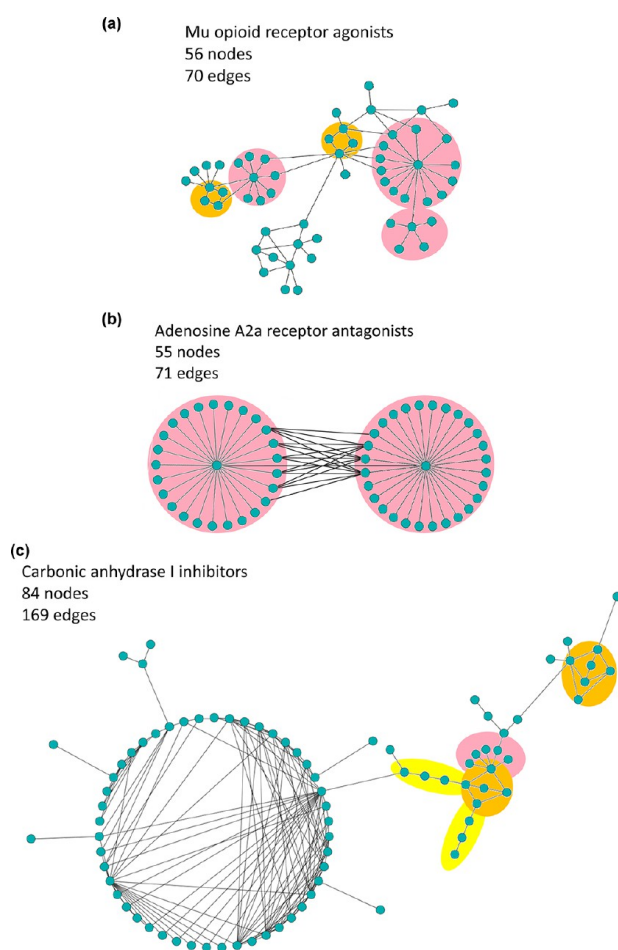


Figure 3. Large activity cliff clusters. Three exemplary large activity clusters with unique topology are shown in detail. These clusters comprise 56 (a), 55 (b), and 84 (c) compounds with different specific activities that form 70, 71, and 169 activity cliffs, respectively. Recurrent topological motifs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

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

Frequency of Occurrence. Table 7 reports the most frequently observed topologies (including isolated cliffs) and cluster size variations. With 335 instances, chains with three compounds represented the most frequent activity cliff clusters, followed by stars with four (122 instances) and five compounds

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

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

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

(70), chains with four (53), and twin stars with five compounds (43). The basic rectangle consisting of four compounds was 26 times observed. Thus, the most frequently occurring activity cliff clusters were of relatively small size. Larger clusters that were also at least three times observed included, among others, stars with 12, twin stars with 11, or modified rectangles with nine compounds.

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

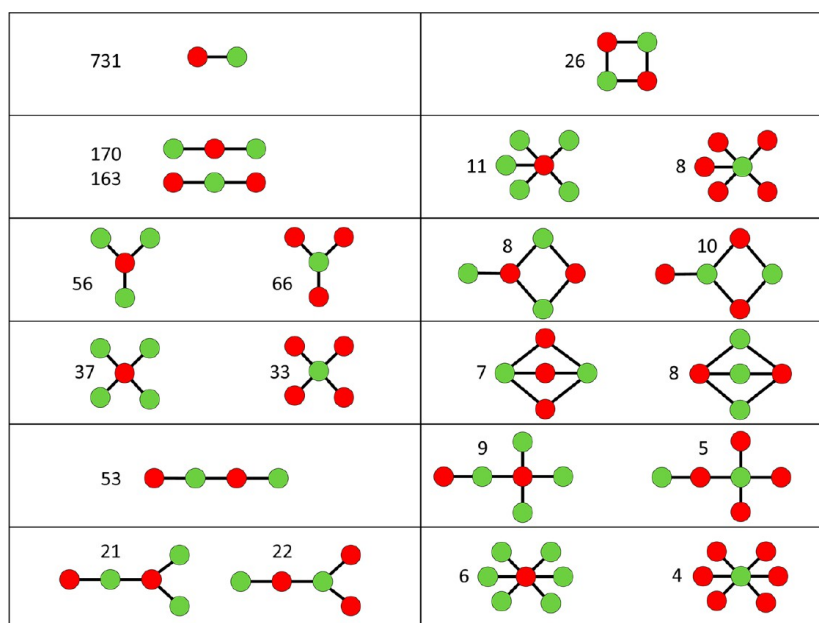


Figure 4. Frequently occurring activity cliff cluster topologies. Shown are the 12 most frequently observed cluster topologies. The number of occurrences is reported. Nodes representing highly and lowly potent cliff partners are colored red and green, respectively. Compounds that were both highly and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

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

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

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

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

Figure 4 shows the 12 most frequently occurring cluster topologies (including isolated cliffs). For nine of the 11 multicliff topologies, different subsets with alternative arrangements of highly and lowly potent activity cliff partners were identified. The exceptions were isolated cliffs, the rectangle, and chain with four compounds each for which no alternative subsets were possible. Small chains with three compounds representing two activity cliffs formed by two highly and one lowly potent cliff partner (170 instances) or one highly potent and two lowly potent partners (163) dominated the topology distribution. In addition, stars with four compounds representing three cliffs or five compounds representing four cliffs were also frequently

observed, with a total of 122 and 70 instances, respectively. These topologies required the combination of a highly potent cliff compound with three or four lowly potent ones or vice versa. Chains with four compounds including two highly and two lowly potent cliff partners that formed three activity cliffs were detected 53 times (compared to 26 instances of the basic rectangle having the same compound composition but forming four cliffs). As shown in Figure 4, topology extensions such as modified and nested rectangles or modified chains were also recurrent.

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

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

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

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

Table 10. Clusters with Largest Numbers of Activity Cliffs^a

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

^aThe top 20 clusters with largest numbers of activity cliffs and their composition are reported. TID stands for target ID.

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

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

inhibitors and three others with different adenosine receptor ligands. Overall, ligands of different G protein coupled receptors formed the majority of large activity cliff clusters.

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

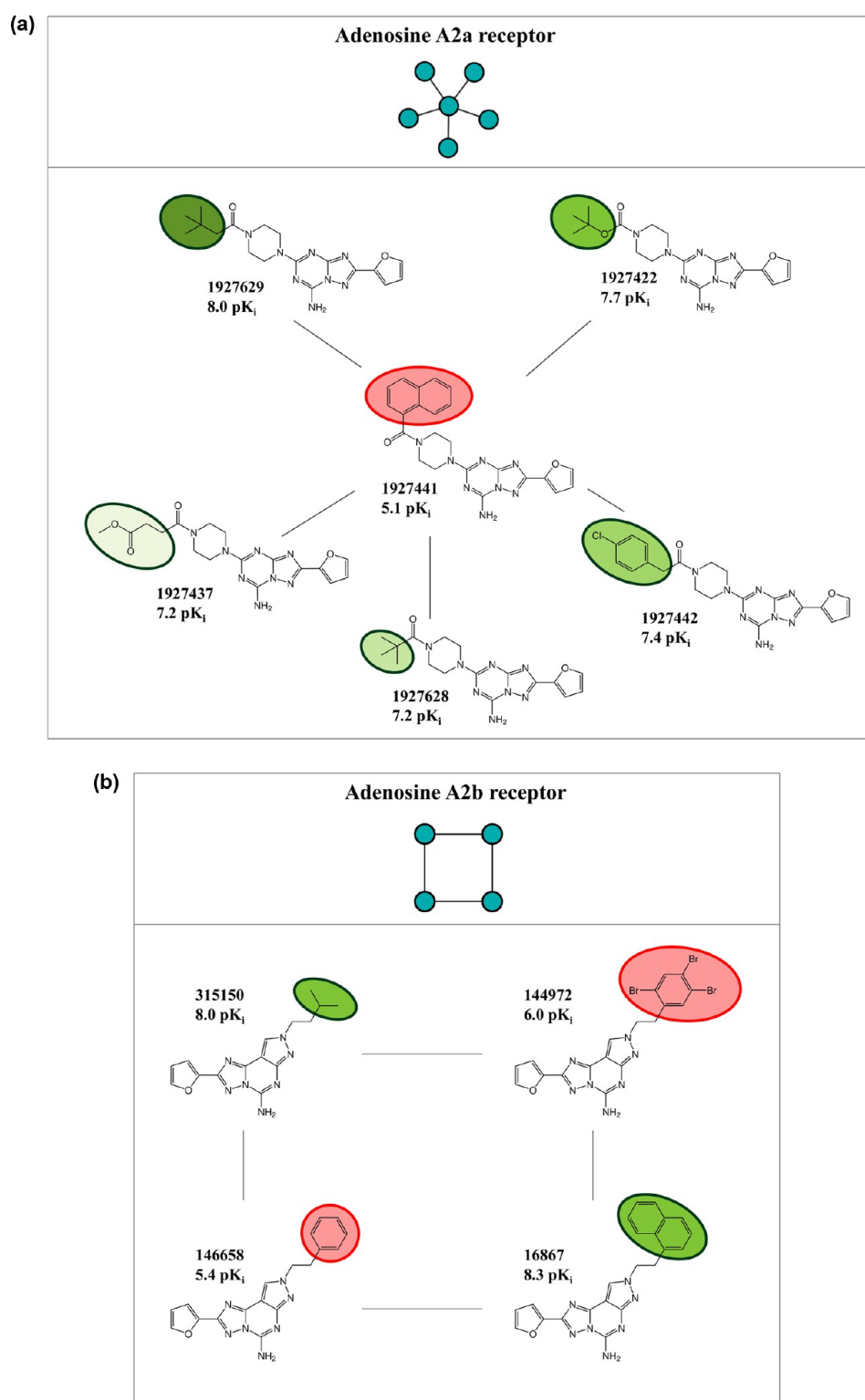


Figure 5. Exemplary cluster topologies and compounds. Shown are compounds forming representative activity cliffs of exemplary cluster topologies (a star, b rectangle). Structural modifications are highlighted and colored according to compound activity (red, low activity; green, high activity). Examples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

advantage that they immediately reveal all compounds participating in a cluster and all activity cliffs comprising the cluster, as illustrated in Figure 5, without the need to search for additional compounds that might further extend given activity

cliff(s). Moreover, a cluster of similar size sharing the same topology can be selected from a given target set and analyzed in a comparative manner, which further increases the amount of SAR information that can be extracted in an organized manner from

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

CONCLUSIONS

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

ASSOCIATED CONTENT

Supporting Information

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

AUTHOR INFORMATION

Corresponding Author

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

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

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

REFERENCES

- (1) Maggiora, G. M. On Outliers and Activity Cliffs – Why QSAR often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535–1535.
- (2) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J. Med. Chem.* **2012**, *55*, 2932–2942.
- (3) Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. *J. Med. Chem.* In press, DOI: 10.1021/jm401120g.
- (4) Hu, Y.; Stumpfe, D.; Bajorath, J. Advancing the Activity Cliff Concept [v1; ref. status: indexed, <http://f1000r.es/1wf>]. *F1000Research* **2013**, *2*, 199 (DOI: 10.12688/f1000research.2-199.v1).
- (5) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity Landscape Representations for Structure-Activity Relationship Analysis. *J. Med. Chem.* **2010**, *53*, 8209–8223.
- (6) Peltason, L.; Iyer, P.; Bajorath, J. Rationalizing Three-Dimensional Activity Landscapes and the Influence of Molecular Representations on Landscape Topology and the Formation of Activity Cliffs. *J. Chem. Inf. Model.* **2010**, *50*, 1021–1033.
- (7) Medina-Franco, J. L.; Martínez-Mayorga, K.; Bender, A.; Marín, R. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. Characterization of Activity Landscapes using 2D and 3D Similarity Methods: Consensus Activity Cliffs. *J. Chem. Inf. Model.* **2009**, *49*, 477–491.
- (8) Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2005; pp 271–285.
- (9) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 339–348.
- (10) Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J. MMP-Cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs. *J. Chem. Inf. Model.* **2012**, *52*, 1138–1145.
- (11) Wassermann, A. M.; Dimova, D.; Bajorath, J. Comprehensive Analysis of Single- and Multi-Target Activity Cliffs Formed by Currently Available Bioactive Compounds. *Chem. Biol. Drug Des.* **2011**, *78*, 224–228.
- (12) Stumpfe, D.; Bajorath, J. Frequency of Occurrence and Potency Range Distribution of Activity Cliffs in Bioactive Compounds. *J. Chem. Inf. Model.* **2012**, *52*, 2348–2353.
- (13) Vogt, M.; Huang, Y.; Bajorath, J. From Activity Cliffs to Activity Ridges: Informative Data Structures for SAR Analysis. *J. Chem. Inf. Model.* **2011**, *51*, 1848–1856.
- (14) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–D1107.
- (15) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A

Summary

The total of 20,080 MMP-cliffs formed by currently available bioactive compounds were organized in a global activity-cliff based network consisting of a total of disjoint 2072 clusters. Most of these cliff communities were of small to moderate size and represented 450 distinct topologies. In addition, 39 topologies were detected that occurred more than three times. Based on frequency analysis, three main topology categories (i.e., star, chain, and rectangle) and a few extensions of these topologies were identified. Chains of limited size represented the most frequent topology. Surprisingly, compounds having both highly and weakly potent partners were only rarely detected. Overall, the majority of the clusters were small and had well-defined topologies indicating that they might display similar SAR characteristics. My contributions to this study have been to aid in the analysis and classification of the activity cliff cluster topologies.

To further extend the activity cliff concept, we have introduced the notion of *promiscuity cliffs* formed by structurally similar compounds having significant differences in the number of target annotations. On the basis of promiscuity cliffs, the relation between chemical structures and compound promiscuity (structure-promiscuity relationship) can be explored.

Chapter 9

Matched Molecular Pair Analysis of Small Molecule Microarray Data Identifies Promiscuity Cliffs and Reveals Molecular Origins of Extreme Compound Promiscuity

Introduction

Compound promiscuity, or the property of a small compound to specifically interact with multiple biological targets, is a major topic in drug development and pharmaceutical research.³ However, the molecular origin of compound promiscuity is currently only little understood and mostly studied on the basis of molecular frameworks. We have further extended the activity cliff concept to explore compound promiscuity from a phenotypic point of view. To these ends, we have utilized the MMP framework to search a recently released small molecule microarray data set comprising 15,252 publicly available compounds

screened against 100 unrelated proteins⁴ for structural modifications that might ultimately change the degree of compound promiscuity.

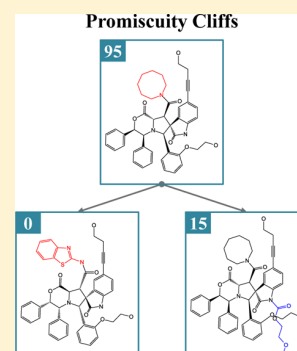
Matched Molecular Pair Analysis of Small Molecule Microarray Data Identifies Promiscuity Cliffs and Reveals Molecular Origins of Extreme Compound Promiscuity

Dilyana Dimova,[†] Ye Hu,[†] and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit, Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113 Bonn, Germany

Supporting Information

ABSTRACT: The study of compound promiscuity is a hot topic in medicinal chemistry and drug discovery research. Promiscuous compounds are increasingly identified, but the molecular basis of promiscuity is currently only little understood. Utilizing the matched molecular pair formalism, we have analyzed patterns of compound promiscuity in a publicly available small molecule microarray data set. On the basis of our analysis, we introduce “promiscuity cliffs” as pairs of structural analogs with single-site substitutions that lead to large-magnitude differences in apparent compound promiscuity involving between 50 and 97 unrelated targets. No substructures or substructure transformations have been detected that are generally responsible for introducing promiscuity. However, within a given structural context, small chemical replacements were found to lead to dramatic promiscuity effects. On the basis of our analysis, promiscuity is not an inherent feature of molecular scaffolds but can be induced by small chemical substitutions. Promiscuity cliffs provide immediate access to such modifications.



INTRODUCTION

Target promiscuity of small molecules is a much investigated topic in medicinal chemistry, for several reasons. First, the binding behavior of a promiscuous compound might be associated with nonspecific binding events, as exemplified by frequent hitters in biological screens.¹ Second, specific interactions of compounds with multiple (related or unrelated) targets might give rise to polypharmacological behavior^{2–5} and also provide a basis for drug repurposing.^{6,7} Third, increasing evidence that many bioactive compounds do act on multiple targets is beginning to change the single-target specificity paradigm that has long governed drug discovery and design efforts.^{8–11} Previous studies have mostly addressed compound promiscuity through database mining,^{5,12,13} for example, by identifying molecular scaffolds that are recurrent in promiscuous compounds,¹² or have focused on polypharmacology by detecting new targets for existing drugs⁵ and by studying side effects.¹³

Most information about compound promiscuity is currently obtained from target annotations of bioactive compounds collected from literature resources and stored in major compound data repositories, such as ChEMBL.¹⁴ In addition, promiscuity information might also be obtained by comparing screening libraries across different bioassays available in PubChem,¹⁵ although this information is limited at present and principally confined to screening hits. Compound promiscuity can experimentally be assessed by systematically testing compound collections on arrays of diverse targets. Unfortunately, such compound profiling data is currently rarely available, at least in the public domain. However, there are a

few notable exceptions. For example, a data set recently released by a group from Abbott Laboratories contains 1473 compounds with reported activities against 1–122 different kinases from a representative sample of the kinome.¹⁶ While this data set provides an excellent test case for large-scale SAR exploration,¹⁷ it is not suitable for promiscuity analysis beyond kinases. Furthermore, Schreiber and colleagues have reported a small molecule microarray experiment that involved screening of diverse compounds against a total of 100 sequence-unrelated targets.¹⁸ The data released as a part of this investigation are highly attractive for a systematic assessment of compound promiscuity. In their original study, Clemons et al. assembled a total of 15 252 compounds from three different sources including compounds commercially available from medicinal chemistry vendors (CCs), natural products (NPs), and compounds originating from diversity-oriented synthesis (DCs).¹⁸ These compounds were then printed on glass slides through surface chemistry or noncovalent absorption and tested against 100 sequence-unrelated soluble proteins. These proteins were selected to represent a total of 145 different InterPro domain classification types.¹⁹ Purified tacked proteins were incubated on microarrays, and proteins bound to array compounds were detected with labeled monoclonal antibodies. These experiments produced a binary readout of activity, i.e., a compound was classified as active against a target or not. Hence, given the nature of microarray experiments, no exact activity measurements were obtained. However, these data

Received: September 7, 2012

Published: October 10, 2012

reflect binding patterns of compounds across a large array of different targets and are thus suitable for the analysis of compound promiscuity or specificity. Clemons et al. determined the distribution of active compounds and analyzed their data primarily considering measures of stereochemical and shape complexity. They found that NPs generally yielded lower hit rates than synthetic compounds and that both NPs and DCs produced many more specific hits than CCs. Increasing stereochemical and shape complexity generally favored compound specificity, as one might anticipate. However, it was also observed that 16% of CCs and 3% of DCs were promiscuous in nature. Clemons et al. found that a spirooxindole moiety was recurrent in the promiscuous subset of DCs. By contrast, possible structural origins of promiscuity among CCs did not become apparent in the course of the analysis. However, a key finding has been that compounds with apparent target selectivity were clearly enriched among DCs compared to CCs.¹⁸

We have been interested in exploring compound promiscuity from a structural perspective, encouraged by the microarray analysis efforts of Schreiber and colleagues involving 100 sequence-unrelated targets. For a thorough structural assessment, we have carried out a matched molecular pair (MMP) analysis²⁰ of all compounds in this data set. We reasoned that MMPs might provide direct access to structural features implicated in promiscuity because compounds forming an MMP are only distinguished by the exchange of a single substructure with limited size. On the basis of our analysis, structural relationships between nonpromiscuous and highly promiscuous compounds were established and substructures were identified that induced large-magnitude promiscuity within a given structural context. MMPs included compounds with very large differences in the number of targets they were active against, leading to the introduction of promiscuity cliffs.

MATERIALS AND METHODS

Compound Data. The publicly released microarray data set¹⁸ contained 15 252 compounds. Each compound was screened against 100 sequence-unrelated proteins. A total of 3433 compounds were active against 1–97 proteins. Compound structures were examined and standardized using the Molecular Operating Environment²¹ and transformed into SMILES strings.²² Compounds with unique SMILES strings were retained. Following these procedures, 15 042 compounds remained for MMP generation including 6151 CCs, 6437 DCs, and 2454 NPs.

Matched Molecular Pair Analysis. An MMP is defined as a pair of compounds that only differ by a structural change at a single site,^{20,23} as illustrated in Figure 1. Compounds forming an MMP are interconverted by the exchange of two substructures, which is termed a chemical transformation.²³ Accordingly, the MMP formalism is descriptor-independent, metric-free, and chemically intuitive. For example, it has been applied to characterize activity cliffs and bioisosteric replacements.^{24–26} MMPs were generated using an in-house implementation of the Hussain and Rea algorithm.²³ Following this approach, conserved core structures and variable substituents of MMPs are stored as keys and values in an index table, respectively. The size of an exchanged substructure (value) was limited to maximally 13 non-hydrogen atoms and the size difference between exchanged substructures to maximally eight non-hydrogen atoms. This was done to restrict the size of exchanged fragments to chemically meaningful replacements.²⁶ In addition, MMP formation was further restricted by the requirement that the core structure of a qualifying compound (key) had to be at least twice the size of each exchanged substructure (value). Application of these size restrictions previously yielded chemically intuitive transformations in an MMP-based study of activity cliffs.²⁶ Furthermore, if several transformations generated the

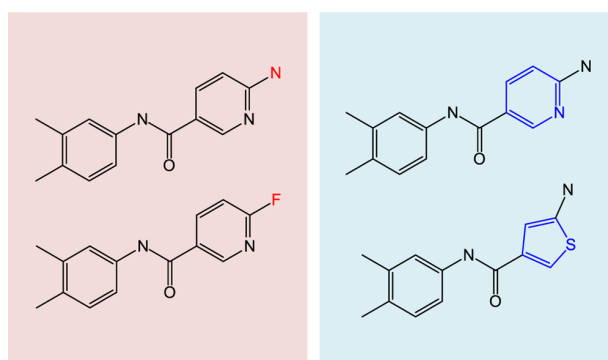


Figure 1. Matched molecular pairs. Two pairs of compounds forming exemplary MMPs are shown. Exchanged fragments are colored in red (left) or blue (right).

same MMP, only the transformation comprising the smallest number of atoms was retained. Following this protocol, MMPs were systematically generated for all 15 042 microarray compounds.

All MMP and data-mining calculations were carried out with in-house generated Java programs or KNIME²⁷ protocols. An MMP-based compound network was drawn with Cytoscape.²⁸

Promiscuity Cliff Criteria. On the basis of our analysis, so-called “promiscuity cliffs” were introduced by applying the following criteria:

- (1) A compound pair formed a transformation size-restricted MMP (as explained above).
- (2) The number of activity annotations of the compounds forming an MMP differed by at least 50 targets, hence indicating large-scale differences in apparent promiscuity.

Accordingly, promiscuity cliffs represented closely related compounds (mostly analogs) with limited structural variations, but large differences in the number of target annotations. These cliffs were systematically explored in the small molecule microarray data set.

RESULTS AND DISCUSSION

MMP Distribution. From the entire compound set, a total of 30 954 nonredundant MMPs were generated that involved a total of 8010 compounds and yielded 7256 different transformations. Most of these transformations were represented by a single MMP or small numbers of MMPs. Differences in the number of target annotations between compounds forming an MMP were evaluated. Therefore, for each MMP, the target profiles of its two compounds were compared. The results are reported in Figure S1 of the Supporting Information. Figure 2 reports the distribution of MMPs over increasing differences in target numbers. Compounds forming 18 251 MMPs (~59%) did not differ in the number of targets they were active against. Only 995 of these MMPs were active against the same number of targets, but different targets (Figure S1 of the Supporting Information). Hence, compounds comprising these 18 251 MMPs displayed the same or comparable levels of promiscuity and were thus of low priority for our analysis.

By contrast, compounds in 829 (~2.7%) and 126 MMPs (~0.4%) differed in their activity by 10 or more and 50 or more targets, respectively, thus revealing structurally similar compounds associated with unexpectedly large differences in apparent promiscuity. As a pinnacle of these trends, 33 MMPs were identified in which compounds differed by 90 or more targets. Taken together, these findings were rather surprising. The 126 MMPs in which activity annotations of compounds differed by 50 or more targets (highlighted in Figure 2) were classified as promiscuity cliffs and subjected to further analysis.

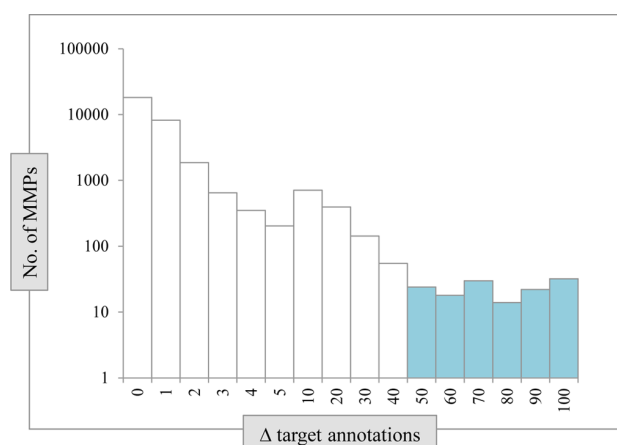


Figure 2. MMPs and target annotations. MMP counts are reported (on a logarithmic scale) for increasing differences in the number of targets MMP-forming compounds were active against. On the horizontal axis, “ Δ target annotations” reports binned differences in target numbers. For example, “1”, “10”, and “100” mean that compounds forming an MMP differed by exactly 1, 6–10, and 91–100 targets, respectively. Sections of the histogram that represent MMPs with a difference of 50 or more targets are highlighted.

Given currently available data, one cannot be certain that binding to a large number of targets might always be specific (in fact, in some instances, this might be unlikely), and which

role local concentration effects on arrays might play. Given that compound promiscuity can have several origins and is influenced by multiple factors, as discussed in the Introduction, the analysis of apparent promiscuity on the basis of compound activity profiles takes these factors implicitly into account. On the basis of the original array data analysis reported by Clemons et al., experimental variances were clearly limited to the level expected for microarrays.

We also identified a total of 1146 MMPs that were formed between an inactive compound and an active compound with at least five target annotations, as reported in Figure S2 of the Supporting Information. These MMPs contained compounds active against 5–95 targets. Of these, 58 MMPs qualified as promiscuity cliffs.

Molecular Properties. For 117 compounds involved in the formation of the 126 promiscuity cliffs, four different physicochemical properties were calculated using the Molecular Operating Environment,²¹ including molecular weight, octanol/water (o/w) partition coefficient ($\log P$), and the numbers of acidic and basic atoms. The distribution of molecular weight is reported in Figure 3a. Compounds that were inactive or active against less than five targets (left region of the plot) covered a broad range, from about 400 to nearly 1000 Da. However, most of the promiscuous compounds (right region) displayed a narrower range. Figure 3b reports the correlation between the changes in molecular weight and promiscuity for individual cliffs. No obvious trends were observed.

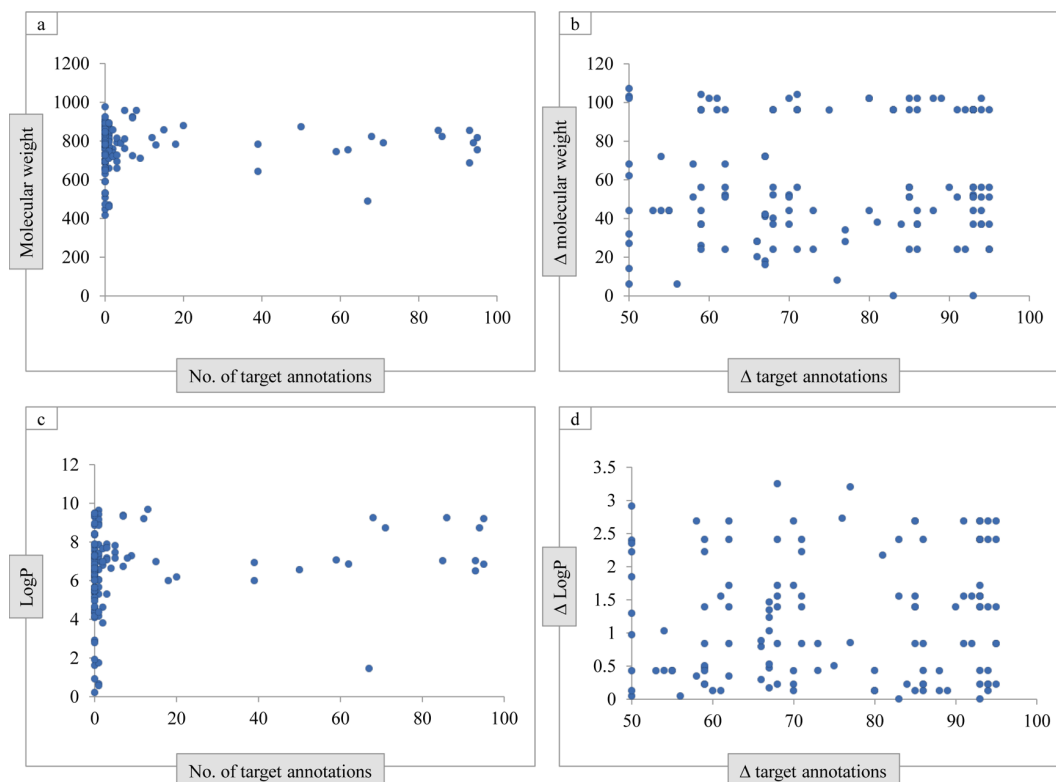
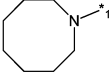

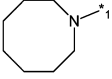
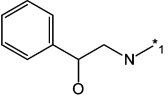
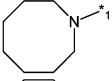
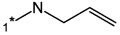
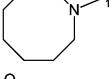
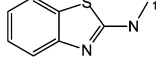
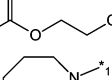

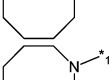

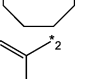
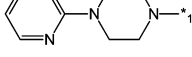
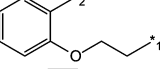

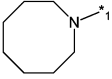

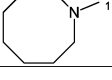



Figure 3. Distribution of molecular properties. For compounds involved in the formation of promiscuity cliffs, the distributions of their molecular weight and o/w partition coefficient ($\log P$) are shown in parts a and c, respectively, as a function of the number of target annotations. In these plots, each dot represents a cliff-forming compound. In addition, for promiscuity cliffs, the distributions of the difference in molecular weight and $\log P$ are shown in parts b and d, respectively, as a function of the difference in the number of target annotations (i.e., differences in the degree of promiscuity). Here, each dot represents a compound pair forming a promiscuity cliff.

Table 1. Ranked Transformations^a

Rank	Transformation	No. of promiscuity cliffs	Total no. of MMPs	Δ target annotations			
				Min	Max	Median	
1			11	42	0	95	6.5
2			11	42	0	95	5
3			11	42	0	95	6
4			11	42	0	95	5
5			11	148	0	94	1
6			10	25	0	93	18
7			10	42	0	95	7
8			7	106	0	94	1
9			6	114	0	88	1
10			4	6	0	93	65

^aThe top 10 transformations most frequently found in promiscuity cliffs are listed. The number of promiscuity cliffs and the total number of MMPs containing each transformation are reported. In addition, the minimal (Min) and maximal (Max) differences in the number of target annotations among MMP-forming compounds and median values are given.

Figure 3c shows the distribution of log *P* values. Analogously to the observations made for molecular weight, inactive and nonpromiscuous compounds also covered a broad range of log *P* values. Most of the promiscuous compounds had a much narrower range, i.e., from 6 to 10. On the other hand, in the area of high lipophilicity (upper region of the plot), both nonpromiscuous and promiscuous compounds were found. Although many promiscuous compounds had relatively high log *P* values (as one might expect), there was no detectable correlation between the changes in lipophilicity and the difference in promiscuity, as shown in Figure 3d.

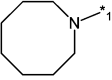
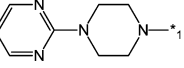
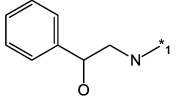
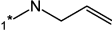
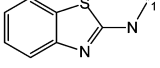
In addition, the protonation states of these compounds were analyzed by counting the numbers of acidic and basic atoms. Nearly all compounds were neutral and only one compound was found to be basic.

Transformations. The 126 MMPs representing promiscuity cliffs encoded 38 unique transformations representing different structural changes (as reported in Table S1 of the Supporting Information). These transformations were ranked according to the number of promiscuity cliffs they occurred in. Table 1 reports the 10 top-ranked transformations. Individual transformations were detected in up to 11 promiscuity cliffs. Notably, eight of the top 10 transformations involved an azocane ring. We calculated the total number of MMPs that represented each of the 38 transformations (including

promiscuity cliffs and others). The results for the top 10 transformations are also reported in Table 1. The total number of MMPs ranged from six to 148. We next determined whether these transformations exclusively occurred in MMPs with large target number differences, i.e., whether they represented promiscuity-inducing transformations. Therefore, target number differences in all MMPs representing a given transformation were analyzed. For the top 10 transformations, the minimal and maximal differences in target numbers between MMP-forming compounds and median values are reported in Table 1. For example, for the top-ranked transformation, the median value was 6.5 and MMPs with no target number differences existed. In three other cases, median values of 1 were obtained. Hence, many promiscuity cliff-containing transformations also occurred in MMPs with small target number differences (or no differences). None of the 38 transformations was found to exclusively occur in promiscuity cliffs or other MMPs with large target number differences. Hence, no chemical transformations were detected that consistently induced large-magnitude compound promiscuity.

Substructures. Following the analysis of transformations, we ranked individual substructures involved in these transformations according to the number of promiscuity cliffs in which they occurred (excluding substructures comprising single atoms). Table 2 shows the top five substructures that were

Table 2. Ranked Substructures^a

Rank	Substructure	No. of promiscuity cliffs	Total no. of MMPs	No. of compounds	No. of annotations		
					Min	Max	Median
1		68	272	86	0	97	6.5
2		14	174	84	0	50	0
3		12	90	1834	0	97	0
4		12	118	625	0	48	0
5		12	90	116	0	28	1.5

^aThe top five substructures most frequently found in promiscuity cliffs are listed. The number of promiscuity cliffs and the total number of MMPs that contain each substructure are reported. In addition, the total number of compounds containing each substructure is reported. Furthermore, the minimal (Min) and maximal (Max) number of target annotations among these compounds and median values are given.

found in more than 10 promiscuity cliffs. Table S2 of the Supporting Information reports all 37 qualifying substructures involved in the formation of cliffs. These substructures were diverse. Corresponding to observations made for transformations, the azocane ring found in 68 promiscuity cliffs was the top-ranked substructure in Table 2. The top five substructures occurred in a total number of 90–272 MMPs and 84–1834 compounds. As also reported in Table 2, the number of targets that compounds containing each substructure were active against greatly varied and also yielded low median values. In three instances, median values of zero were obtained, indicating that at least half of the compounds containing a highly ranked substructure were inactive. As expected on the basis of our transformation analysis, no substructure was found to exclusively occur in promiscuous compounds.

Promiscuous Compounds. All 117 compounds involved in the formation of the 126 promiscuity cliffs were used to generate a molecular network in which nodes represented compounds and edges promiscuity cliffs, as shown in Figure 4a. In this network representation, a number of “promiscuity hubs” became apparent, i.e., compounds with a large number of target annotations involved in the formation of multiple cliffs. It should be noted that these compounds were not only highly promiscuous, but also could be transformed into multiple compounds with limited or no promiscuity through small chemical modifications. The five most prominent promiscuity hubs are highlighted in Figure 4a. These hubs were active against more than 90 targets each and involved in the formation of 9–11 promiscuity cliffs. Their structures are shown in Figure 4b. A characteristic feature of all five compounds was that they contained both the azocane ring and spirooxindole rings (the latter identified by Schreiber and colleagues¹⁸ as a single promiscuity marker in DCs; vide supra). Because of the very large number of targets that promiscuity hubs were active against, it is conceivable that they might at least in part also engage in nonspecific interactions (vide supra).

Promiscuity Cliffs. The co-occurrence of the azocane and spirooxindole substructures in many highly promiscuous compounds suggested the possibility that combinations of substructures (rather than individual ones) might be promiscuity determinants. This possibility could be directly explored because the hubs we identified participated in the formation of multiple promiscuity cliffs. Figure 5 shows examples of prominent promiscuity cliffs containing the azocane and spirooxindole substructures (additional examples are provided in Figure S3 of the Supporting Information). Comparison of cliff-forming compounds clearly revealed that co-occurrence of the azocane and spirooxindole moieties was not a major promiscuity determinant. In the promiscuity cliffs in Figure 5a,b, removal of the azocane ring rendered highly promiscuous compounds (with activity against 95 and 94 targets, respectively) inactive. All compounds in these cliffs also contained the spirooxindole moiety. The cliff forming compounds in Figure 5c both contained the azocane and spirooxindole rings. However, a change in the position of an aliphatic substituent from the para to ortho in the phenyl ring at the lower right was sufficient to transform a highly promiscuous compound into an inactive one. In Figure 5d, the compound containing the para-substituted phenyl ring was also highly promiscuous (i.e., active against 93 targets), whereas the presence of a hydroxyl group at the same position dramatically reduced promiscuity to five targets. However, the ortho-substituted phenyl ring in a different structural context, shown in Figure 5e, was highly promiscuous in contrast to the corresponding analog in Figure 5c. Moreover, compounds containing the para-substituted phenyl ring but different substitutions at the spirooxindole moiety displayed very different degrees of promiscuity (Figure 5e). Taken together, these comparisons revealed a strong structural context dependence of chemical modifications, leading to the formation of promiscuity cliffs. There was no individual substructure or transformation that consistently caused large-magnitude

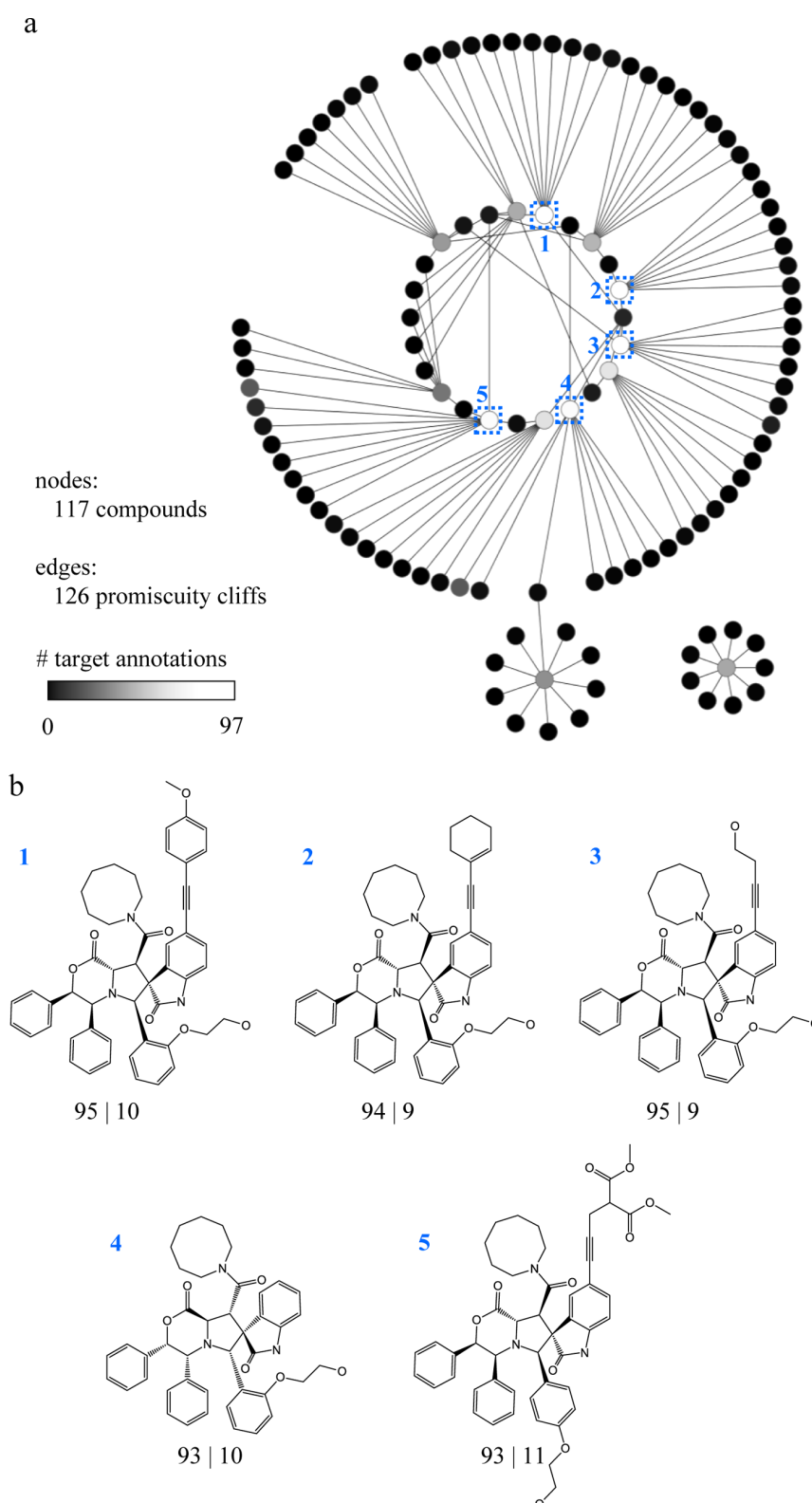


Figure 4. Promiscuity cliff network. (a) MMP-based compound network focusing on promiscuity cliffs. Nodes represent compounds, and edges indicate promiscuity cliffs. Nodes are gray-scaled according to the number of target annotations using a continuous spectrum from black (0 targets; inactive) to white (97 targets; most promiscuous). Five highly promiscuous compounds that were active against more than 90 targets and involved in the formation of 9–11 cliffs are boxed and numbered. Their structures are shown in part b. For each compound, the number of targets it was active against and the number of cliffs it was involved in are reported. For example, “95 | 10” means that the compound was active against 95 targets and involved in the formation of 10 promiscuity cliffs.

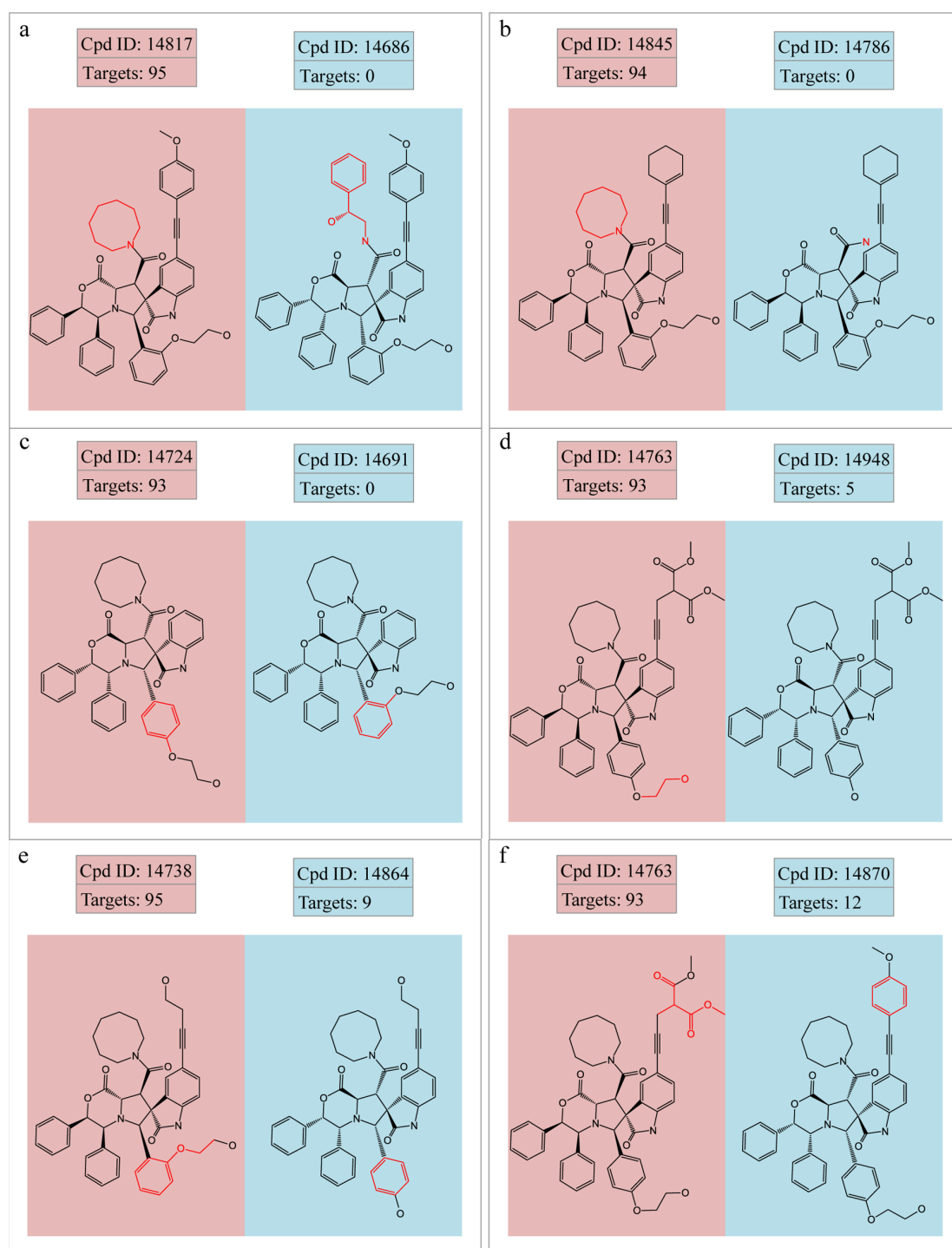


Figure 5. Promiscuity cliffs. Shown are representative MMPs in which activity annotations of compounds differed by more than 80 targets. For each compound, the compound ID and the number of targets it was active against are reported. The promiscuous compound of each cliff is shown on the left and the exchanged fragments are colored red.

promiscuity effects and exclusively occurred in promiscuous compounds.

What Do We Learn about Promiscuity from a Medicinal Chemistry Perspective? On the basis of the data available to us, it is not possible to conclude with certainty to what extent highly promiscuous compounds engage in specific and/or nonspecific interactions with targets. It is of

course unlikely that a compound might form specific interactions with 90 or more diverse targets, even if the interactions were clearly detectable under the given experimental conditions. Hence, it is appropriate to consider promiscuity from a phenotypic point of view in the context of our analysis, given the requirement to analyze the data at face value and avoid overinterpretation. However, it should be

noted that only a small fraction of the array compounds were promiscuous in nature and that the formation of promiscuity cliffs was a rare event, thus indicating that the microarray data were suitable for a systematic analysis of promiscuity effects. As we have shown, only a small fraction of MMPs generated from the entire microarray data set combined compounds with notable differences in the number of targets they were active against. Taking this into account, the detection of cliffs in which structurally similar compounds differed in their activity by 50 or more targets is considered a striking finding, regardless of underlying molecular mechanisms.

For medicinal chemistry, a number of findings reported herein are of immediate relevance. It is evident that the MMP-based approach provides a direct and chemically intuitive access to small structural modifications, leading to large-magnitude promiscuity effects. Previously, a number of structural frameworks have been identified that were highly recurrent in promiscuous compounds across different target families.¹² However, it has remained largely unclear from a medicinal chemistry perspective thus far whether certain molecular frameworks carry an intrinsic likelihood of promiscuity and/or might have frequent hitter character. After all, promiscuity is determined for compounds, not their frameworks. Importantly, the findings presented herein do not promote a framework-centric view of promiscuity. Thus, for the evaluation and prioritization of compound series for medicinal chemistry, frameworks should not primarily be considered as an intrinsic source of promiscuity and potential lack of compound specificity. Rather, we demonstrate that small chemical modifications can trigger large-magnitude promiscuity effects. Importantly, these effects depend on the specific structural environment in which these modifications occur. On the basis of our analysis, substitutions that induce promiscuity in any structural environment were not identified. Thus, in medicinal chemistry, it is important to evaluate promiscuity for individual compounds in series that are preferred from an SAR perspective; observed specificity of certain analogs within a series does not guarantee that others are not highly promiscuous. Taken together, these findings further extend our view of molecular origins of promiscuity, putting strong emphasis on the context-dependence of promiscuity-inducing structural modifications. The analysis of compounds in cliff forming MMPs provided a focal point for the identification of such chemical changes that might have otherwise not been detected.

CONCLUSIONS

Herein, we have analyzed compound promiscuity on the basis of small molecule microarray data involving ~15 000 compounds and 100 sequence-unrelated targets. These microarray data provide a binary readout of compound activity and are likely influenced, for example, by variance and local concentration effects associated with printing of compounds on solid surfaces by different mechanisms. Nevertheless, as clearly indicated by the results of Clemons et al., who conducted the microarray experiments, the data revealed meaningful binding patterns and systematic trends concerning compound selectivity and, as demonstrated in our study, promiscuity. In the current analysis, we have focused on identifying closely related compounds with large difference in promiscuity, leading to the introduction of promiscuity cliffs. From these compound pairs, chemical modifications at individual sites have become apparent that led to promiscuous binding behavior. Chemical changes

were identified that caused large-magnitude promiscuity effects. We have shown that no individual substructure or transformation involved in these effects exclusively occurred in promiscuous compounds. Rather, they were distributed across compounds with different levels of promiscuity or no apparent promiscuity. On the basis of currently available data, promiscuity is not an inherent feature of certain structural frameworks. However, we have shown that chemical modifications could trigger promiscuity within specific structural contexts. Exemplary promiscuity cliffs have revealed that similar substitutions in different structural environments can lead to promiscuity effects of different magnitude, or even opposite effects (i.e., increase vs reduction in target numbers). On the basis of our analysis, small structural modifications of nonpromiscuous compounds can lead to substantial promiscuity. However, these effects are structural-context-dependent.

ASSOCIATED CONTENT

Supporting Information

Figure S1 reports the comparison of target annotations for compound pairs forming MMPs, Figure S2 shows the distribution of the number of targets for compounds that were active against at least five targets and formed MMPs with inactive ones, Figure S3 shows further examples of promiscuity cliffs, and Tables S1 and S2 report transformations encoded by promiscuity cliffs and substructures involved in these transformations, respectively. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

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

Author Contributions

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

Notes

The authors declare no competing financial interest.

ABBREVIATIONS USED

CC, commercial compound; DC, compound from diversity-oriented synthesis; MMP, matched molecular pair; NP, natural product; SAR, structure–activity relationship

REFERENCES

- (1) Feng, B. Y.; Shelat, A.; Doman, T. N.; Guy, R. K.; Shoichet, B. K. High-Throughput Assays for Promiscuous Inhibitors. *Nat. Chem. Biol.* **2005**, *1*, 146–148.
- (2) Paolini, G. V.; Shapland, R. H. B.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global Mapping of Pharmacological Space. *Nat. Biotechnol.* **2006**, *24*, 805–815.
- (3) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating Protein Pharmacology by Ligand Chemistry. *Nat. Biotechnol.* **2007**, *25*, 197–206.
- (4) Hopkins, A. L. Network Pharmacology: The Next Paradigm in Drug Discovery. *Nat. Chem. Biol.* **2008**, *4*, 682–690.
- (5) Keiser, M. J.; Setola, V.; Laggner, C.; Abbas, A. I.; Hufeisen, S. J.; Jensen, N. H.; Kuijter, M. B.; Matos, R. C.; Tran, T. B.; Whaley, R.; Glennon, R. A.; Hert, J.; Thomas, K. L.; Edwards, D. D.; Shoichet, B. K.; Roth, B. L. Predicting New Molecular Targets for Known Drugs. *Nature* **2009**, *462*, 175–181.
- (6) Ashburn, T. T.; Thor, K. B. Drug Repositioning: Identifying and Developing New Uses for Existing Drugs. *Nat. Rev. Drug Discovery* **2004**, *3*, 673–683.

- (7) Chong, C. R.; Sullivan, D. J. New Uses for Old Drugs. *Nature* **2007**, *448*, 645–646.
- (8) Mestres, J.; Gregori-Puigjané, E. Conciliating Binding Efficiency and Polypharmacology. *Trends Pharmacol. Sci.* **2009**, *30*, 470–474.
- (9) Merino, A.; Bronowska, A. K.; Jackson, D. B.; Cahill, D. J. Drug Profiling: Knowing Where It Hits. *Drug Discovery Today* **2010**, *15*, 749–756.
- (10) Koutsoukas, A.; Simms, B.; Kirchmair, J.; Bond, P. J.; Whitmore, A. V.; Zimmer, S.; Young, M. P.; Jenkins, J. L.; Glick, M.; Glen, R. C.; Bender, A. From in Silico Target Prediction to Multi-Target Drug Design: Current Databases, Methods and Applications. *J. Proteomics* **2011**, *74*, 2554–2574.
- (11) Xie, L.; Xie, L.; Kinnings, S. L.; Bourne, P. E. Novel Computational Approaches to Polypharmacology as a Means to Define Responses to Individual Drugs. *Annu. Rev. Pharmacol. Toxicol.* **2012**, *52*, 361–379.
- (12) Hu, Y.; Bajorath, J. Polypharmacology Directed Compound Data Mining: Identification of Promiscuous Chemotypes with Different Activity Profiles and Comparison to Approved Drugs. *J. Chem. Inf. Model.* **2010**, *50*, 2112–2118.
- (13) Campillos, M.; Kuhn, M.; Gavin, A. C.; Jensen, L. J.; Bork, P. Drug Target Identification Using Side-Effect Similarity. *Science* **2008**, *321*, 263–266.
- (14) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–D1107.
- (15) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B. A.; Bolton, E.; Gindulyte, A.; Bryant, S. H. PubChem's Bioassay Database. *Nucleic Acids Res.* **2012**, *40*, D400–D412.
- (16) Metz, J. T.; Johnson, E. F.; Soni, N. B.; Merta, P. J.; Kifle, L.; Hajduk, P. J. Navigating the Kinome. *Nat. Chem. Biol.* **2011**, *7*, 200–202.
- (17) Iyer, P.; Dimova, D.; Vogt, M.; Bajorath, J. Navigating High-Dimensional Activity Landscapes: Design and Application of the Ligand–Target Differentiation Map. *J. Chem. Inf. Model.* **2012**, *52*, 1962–1969.
- (18) Clemons, P. A.; Bodycombe, N. E.; Carrinski, H. A.; Wilson, J. A.; Shamji, A. F.; Wagner, B. K.; Koehler, A. N.; Schreiber, S. L. Small Molecules of Different Origins Have Distinct Distributions of Structural Complexity that Correlate with Protein-Binding Profiles. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 18787–18792.
- (19) *Interpro Sequence Analysis & Classification*, www.ebi.ac.uk/interpro/ (Accessed August 14, 2012).
- (20) Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Cheminformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2004; pp 271–285.
- (21) *Molecular Operating Environment (MOE)*, 2011.10; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2011.
- (22) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
- (23) 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.
- (24) Wassermann, A. M.; Bajorath, J. Chemical Substitutions That Introduce Activity Cliffs across Different Compound Classes and Biological Targets. *J. Chem. Inf. Model.* **2010**, *50*, 1248–1256.
- (25) Wassermann, A. M.; Bajorath, J. Large-Scale Exploration of Biososteric Replacements on the Basis of Matched Molecular Pairs. *Future Med. Chem.* **2011**, *3*, 425–436.
- (26) 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.
- (27) Tiwaria, A.; Sekhar, A. K. T. Workflow Based Framework for Life Science Informatics. *Comput. Biol. Chem.* **2007**, *31*, 305–319.
- (28) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13*, 2498–2504.

Summary

A total of 30,954 MMPs were systematically generated for the set of 15,252 microarray compounds screened against 100 unrelated targets. In addition, 126 MMPs formed by 117 compounds were identified that differed by 50 or more target annotations and qualified as promiscuity cliffs. These 126 cliffs encoded a total of 38 transformations involving 37 distinct substructures. Our findings revealed that no single chemical transformation or a substructure existed that exclusively occurred in promiscuous compounds. In addition, promiscuity hubs were detected that were represented by highly promiscuous compounds involved in multiple promiscuity cliffs. A key observation was that these promiscuity hubs contained both the azocane and spirooxindole rings, suggesting that combinations of substructures (rather than individual ones) might be promiscuity determinants. Taken together, there was a strong context-dependence of promiscuity-inducing chemical modifications. My contributions to the work presented herein include the MMP and promiscuity cliff analysis.

Chapter 10

Conclusions

The analysis of SARs of small bioactive compounds is a central task in medicinal chemistry. However, the multi-faceted nature of SARs and the rapidly growing molecular data greatly challenge SAR exploration. A variety of computational approaches have been introduced thus far to aid in the systematic SAR analysis, with activity landscapes being popular and widely used computational models. Furthermore, the concept of activity cliffs, the cardinal features of activity landscape representations, provides a direct access to SAR determinants in active compounds. Therefore, the major objectives of this dissertation have been the development of novel activity landscape representations and the systematic exploration of activity cliffs in public domain compound repositories. A number of representative studies have been presented.

First, a newly designed activity landscape was introduced for compounds active against multiple targets. This method was based on a numerical encoding scheme of activity profiles of compounds. On the basis of this methodology, compounds forming multi-target activity cliffs could be easily identified and the contribution of individual compounds to global multi-target SARs monitored (Chapter 2). This method was applied to sets of compounds active against a limited number of targets. Furthermore, the LTD map was introduced to navigate high-dimensional activity spaces defined by compounds active against 50 or more targets (Chapter 3). A key feature of the LTD map was that it greatly reduced the complexity of high-dimensional activity spaces by account-

ing for pairwise potency differences between compounds. Furthermore, pairwise activity profile comparisons were made graphically assessable.

The analysis of activity cliffs is of major importance for medicinal chemistry. However, thus far, their study has mostly focused on individual sets of compounds active against a given target. Therefore, a systematic survey was carried out to account for the global distribution of single- and multi-target cliffs across different targets. The results revealed that the majority of activity cliffs were single-target cliffs, yet instances of multi-target cliffs were also detected. In addition, the propensity of active compounds to form activity cliffs was determined (Chapter 4).

In a large-scale SAR profiling experiment, fingerprint-dependent changes in SAR information were determined. For approximately 70% of the test compounds changes in the local SAR discontinuity score were observed when alternative fingerprint representations were used. Further, nearly 30% of the compounds shifted their SAR phenotype. Hence, SAR characteristics were often highly fingerprint-dependent (Chapter 5).

The remainder of the thesis was dedicated to the concept of activity cliffs and their utility to aid in the practical medicinal chemistry. First, the question was investigated whether there was a detectable SAR advantage associated with exploring activity cliffs. To these ends, a computational compound pathway model was designed to represent compound series with steadily increasing potency ultimately leading to highly potent data set compound. Different pathway categories were introduced on the basis of their origin. It was demonstrated that the majority of the most potent data set compounds were reached by pathways. In addition, activity cliff-dependent pathways reached potent data set compounds with a higher relative frequency than activity-cliff independent pathways, clearly indicating an SAR information gain associated with activity cliffs (Chapter 6).

The concept of SAR progression was developed to mimic the exploration of the chemical neighborhood of cliff-forming compounds and indicate to what extent activity cliffs were utilized as starting points for compound optimization efforts (Chapter 7). For the majority of activity cliffs, no structural analogs of the highly potent cliff-forming compounds were reported and hence there was no

CONCLUSIONS

evidence for further chemical exploration. On the other hand, for every sixth to seventh cliff, instances of activity cliff progression were detected that resulted in a significant potency increase of 1 or 2 orders of magnitude. The large fraction of as of yet unexplored chemical neighborhoods indicated that there are many opportunities for further exploration and exploitation of activity cliff in practical medicinal chemistry.

To further complement the activity cliff analysis, the composition and topology of activity cliffs formed by currently available bioactive compounds were thoroughly examined (Chapter 8). The majority of the cliffs were coordinated, i.e., their formation involved multiple active compounds and cliffs. All distinct topologies were systematically determined and recurrent topologies classified. The majority of the cliff communities were of small size and had well-defined topologies, which might be indication of similar SAR patterns present in these communities.

Finally, the concept of activity cliffs was extended to account for structure-promiscuity relationships. By introducing the concept of promiscuity cliffs, promiscuity-inducing structural modifications were identified that significantly altered the degree of compound promiscuity (Chapter 9).

In conclusion, in this thesis computational methods were introduced designed to assist in large-scale SAR exploration. Furthermore, systematic data mining efforts were reported to provide a detailed and further refined view of the activity cliff concept and its utility to aid in the practice of medicinal chemistry.

Additional References

- [1] Metz, J. T.; Johnson, E. F.; Soni, N. B.; Merta, P. J.; Kifle, L.; Hajduk, P. J. Navigating the Kinome. *Nature Chemical Biology* **2011**, *7*, 200–202.
- [2] Hu, Y.; Stumpfe, D.; Bajorath, J. Advancing the Activity Cliff Concept [v1; ref status: indexed, <http://f1000r.es/1wff>]. *F1000Research* **2013**, *2*, 199.
- [3] Hu, Y.; Bajorath, J. High-Resolution View of Compound Promiscuity [v2; ref status: indexed, <http://f1000r.es/1ig>]. *F1000Research* **2013**, *2*, 144.
- [4] Clemons, P. A.; Bodycombe, N. E.; Carrinski, H. A.; Wilson, J. A.; Shamji, A. F.; Wagner, B. K.; Koehler, A. N.; Schreiber, S. L. Small Molecules of Different Origins have Distinct Distributions of Structural Complexity that Correlate with Protein-Binding Profiles. *Proceedings of the National Academy of Sciences* **2010**, *107*, 18787–18792.

Additional Publications

- [5] Kayastha, S.; Dimova, D.; Iyer, P.; Vogt, M.; Bajorath, J. Large-Scale Assessment of Activity Landscape Feature Probabilities of Bioactive Compounds. *Journal of Chemical Information and Modeling*, DOI: 10.1021/ci400677b, in press.
- [4] Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. *Journal of Medicinal Chemistry* **2014**, *57*, 18–28.
- [3] Dimova, D.; Iyer, P.; Vogt, M.; Totzke, F.; Kubbutat, M. H.; Schächtele, C.; Laufer, S.; Bajorath, J. Assessing the Target Differentiation Potential of Imidazole-Based Protein Kinase Inhibitors. *Journal of Medicinal Chemistry* **2012**, *55*, 11067–11071.
- [2] Wassermann, A. M.; Dimova, D.; Iyer, P.; Bajorath, J. Advances in Computational Medicinal Chemistry: Matched Molecular Pair Analysis. *Drug Development Research* **2012**, *73*, 518–527.
- [1] Dimova, D.; Bajorath, J. Computational Chemical Biology: Identification of Small Molecular Probes that Discriminate between Members of Target Protein Families. *Chemical Biology & Drug Design* **2012**, *79*, 369–375.

Eidesstattliche Erklärung

An Eides statt versichere ich hiermit, dass ich die Dissertation “Computational Methods Generating High-Resolution Views of Complex Structure-Activity Relationships” selbst und ohne jede unerlaubte Hilfe angefertigt habe, dass diese oder eine ähnliche Arbeit noch an keiner anderen Stelle als Dissertation eingereicht worden ist und dass sie an den nächstehend aufgeführten Stellen auszugsweise veröffentlicht worden ist:

- [1] Dimova, D.; Wawer, M.; Wassermann, A. M.; Bajorath, J. Design of Multitarget Activity Landscapes that Capture Hierarchical Activity Cliff Distributions. *Journal of Chemical Information and Modeling* **2011**, *51*, 258–266.
- [2] Iyer, P.; Dimova, D.; Vogt, M.; Bajorath, J. Navigating High-Dimensional Activity Landscapes: Design and Application of the Ligand-Target Differentiation Map. *Journal of Chemical Information and Modeling* **2012**, *52*, 1962–1969.
- [3] Wassermann, A. M.; Dimova, D.; Bajorath, J. Comprehensive Analysis of Single-and Multi-Target Activity Cliffs Formed by Currently Available Bioactive Compounds. *Chemical Biology & Drug Design* **2011**, *78*, 224–228.
- [4] Dimova, D.; Stumpfe, D.; Bajorath, J. Quantifying the Fingerprint Descriptor Dependence of Structure-Activity Relationship Information on a

- Large Scale. *Journal of Chemical Information and Modeling* **2013**, *53*, 2275–2281.
- [5] Stumpfe, D.; Dimova, D.; Heikamp, K.; Bajorath, J. Compound Pathway Model To Capture SAR Progression: Comparison of Activity Cliff-Dependent and-Independent Pathways. *Journal of Chemical Information and Modeling* **2013**, *53*, 1067–1072.
- [6] 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.
- [7] Stumpfe, D.; Dimova, D.; Bajorath, J. Composition and Topology of Activity Cliff Clusters Formed by Bioactive Compounds. *Journal of Chemical Information and Modeling*, DOI: 10.1021/ci400728r, in press.
- [8] Dimova, D.; Hu, Y.; Bajorath, J. Matched Molecular Pair Analysis of Small Molecule Microarray Data Identifies Promiscuity Cliffs and Reveals Molecular Origins of Extreme Compound Promiscuity. *Journal of Medicinal Chemistry* **2012**, *55*, 10220–10228.

Dilyana Dimova

Bonn, 2014