# Multi-faceted Structure-Activity Relationship Analysis Using Graphical Representations

Kumulative Dissertation

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

der Mathematisch-Naturwissenschaftlichen Fakultät

der Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Preeti Ramesh Iyer

aus Chennai, Indien

Bonn

October, 2013

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: 16 January, 2014
Erscheinungsjahr: 2014

## Abstract

A core focus in medicinal chemistry is the interpretation of structure-activity relationships (SARs) of small molecules. SAR analysis is typically carried out on a case-by-case basis for compound sets that share activity against a given target. Although SAR investigations are not a priori dependent on computational approaches, limitations imposed by steady rise in activity information have necessitated the use of such methodologies. Moreover, understanding SARs in multi-target space is extremely difficult. Conceptually different computational approaches are reported in this thesis for graphical SAR analysis in single- as well as multi-target space. Activity landscape models are often used to describe the underlying SAR characteristics of compound sets. Theoretical activity landscapes that are reminiscent of topological maps intuitively represent distributions of pair-wise similarity and potency difference information as three-dimensional surfaces. These models provide easy access to identification of various SAR features. Therefore, such landscapes for actual data sets are generated and compared with graph-based representations. Existing graphical data structures are adapted to include mechanism of action information for receptor ligands to facilitate simultaneous SAR and mechanism-related analyses with the objective of identifying structural modifications responsible for switching molecular mechanisms of action. Typically, SAR analysis focuses on systematic pair-wise relationships of compound similarity and potency differences. Therefore, an approach is reported to calculate SAR feature probabilities on the basis of these pair-wise relationships for individual compounds in a ligand set. The consequent expansion of feature categories improves the analysis of local SAR environments. Graphical representations are designed to avoid a dependence on preconceived SAR models. Such representations are suitable for systematic large-scale SAR exploration. Methods for the navigation of SARs in multi-target space using simple and interpretable data structures are introduced. In summary, multi-faceted SAR analysis aided by computational means forms the primary objective of this dissertation.

*For my family*

## Acknowledgments

First and foremost, I would like to express my sincere thanks to my supervisor Prof. Dr. Jürgen Bajorath for his invaluable guidance, continued support, infinite patience and immense encouragement during the course of my PhD study. I would also like to thank Prof. Dr. Michael Gütschow for taking time to review my dissertation as co-referee.

I would like to express my heartfelt gratitude to all my colleagues in the Life Science Informatics group for providing a friendly, interactive and lively working atmosphere. I am especially thankful to Dr. Anne Mai Wassermann, Dr. Dagmar Stumpfe, Dr. Lisa Peltason, Dr. Martin Vogt, Dr. Mathias Wawer, Dr. Vigneshwaran Namasivayam and Dr. Ye Hu for helpful discussions, pleasant and productive collaborations. I also extend my thanks to Dilyana Dimova for collaborative and fruitful discussions.

I also thank the Sonderforschungsbereich (SFB) 704 of the Deutsche Forschungsgemeinschaft for support and funding.

Finally, I would like to express my love and gratitude to my family for their support and understanding during the course of my studies.

# Contents

# Introduction

The development of compounds that specifically interact with given biological targets is the central aspect of medicinal chemistry research. It is often assumed that the chemical structures of these compounds determine their bioactivity. The study of structure-activity relationships (SARs) is largely (but not exclusively) based upon this premise. Furthermore, in accordance with an intuitive postulate, the *similarity-property principle* (SPP), one can also extrapolate that compounds having similar chemical structures would most likely have similar biological activities [1]. Consequently, minor modifications of the chemical structure of an active compound would alter its activity only within a narrow range. However, such straightforward assumptions are not always valid. In many cases, simple structural modifications in a molecule are accompanied by large changes in biological activity, either by dramatically increasing its existing activity or rendering it inactive [2]. Furthermore, despite being structurally related, active compounds may interact differently with their targets [3]. Thus, determining the underlying SARs of bioactive compounds remains a significant challenge in medicinal chemistry.

## Computational Chemical Space and Similarity

Computational approaches are often favored while investigating SARs on a large-scale as systematic comparisons of molecular structure and activity become exceedingly difficult. Such analyses often require a computationally accessible representation for molecular structures and a reference framework that allows their comparison [4]. Mathematical formulations that encode physical and chemical properties of active compounds, known as molecular descriptors, are commonly used molecular representations. A chemical reference space, defined using a set of molecular descriptors, wherein each descriptor constitutes

a dimension, would correspond to a *coordinate-based chemical space*. Thus, compounds projected in such a chemical space would be represented by vectors of their respective descriptor values [4]. Molecules that are structurally similar would ideally be located in close proximity within this space, whereas increasing distances between molecular positions would account for dissimilar compounds. Therefore, construction of meaningful chemical reference spaces is crucial to similarity assessment, and the selection of activity-relevant descriptors is a major challenge [5].

A plethora of descriptors are available as molecular representations [6, 7]. Molecular fingerprints, a popular type of molecular representation, are bitstrings that encode the chemical structure and properties of the compounds [4]. Such fingerprints usually are binary in nature and the bits indicate the presence or absence of specific structural features. Depending on how these features are determined, the resulting fingerprints may vary in their size and complexity. For instance, fragment-based fingerprints like MACCS [8] are generated from a set of predefined structural features. Furthermore, atom environment [9] and extended connectivity [10] fingerprints are derived from all connectivity pathways of specified lengths that exist in a given molecule. Moreover, fingerprints may also be designed to capture possible pharmacophore elements within compounds [11]. Therefore, different types of fingerprints resolve molecular structure at various levels [5].

In addition to generating computationally accessible molecular representations, one must also consider ways to compare them in a quantitative manner and assess the similarity or distance between these representations (and consequently the molecules) within the chemical reference space. However, the concept of similarity in general is representation-dependent [12]. Besides, development of methods that quantify the degree of similarity or dissimilarity between compounds is also required. Many such similarity and distance measures have been reported [6].

Medicinal chemists often need to establish chemically interpretable trends during exploration of SARs. Identifying structural determinants of activity using molecular descriptors or fingerprints is often difficult. The concept of matched molecular pairs (MMPs) has become popular as it provides a frame-

work for studying the structural relatedness among bioactive compounds on a large scale [13]. An MMP consists of a pair of compounds that can be interconverted by a well-defined structural modification restricted to a single site. In addition to single-point MMPs, multi-point MMPs with changes at more than one site have also been defined [14]. Given that the primary objective of MMP analyses is to account for all possible MMPs for given sets of compounds, several algorithms have been reported for such pairwise molecular comparisons.

Two widely employed methodologies include maximum common substructure (MCS) based and systematic molecular fragmentation approaches [13]. A popular fragmentation scheme reported by Hussain and Rea produces molecular fragments through systematic deletion of up to three acyclic single bonds resulting in single, double and triple cuts. Bond deletions result in larger common substructures and smaller transformations. Each larger substructure fragment and the corresponding transformation is indexed as a key and value pair, respectively. The value fragments may have one (single cuts) or more (double and triple cuts) attachment points [15]. Initially, the MMP concept was applied to analyze bioisosteric replacements within drugs and drug-like compounds that conserved the activity against their targets [16]. Several unique bioisosteric transformations have been identified after systematic examination of MMPs formed within compound sets active against different targets and target families obtained from public repositories [17, 18]. Molecular transformations that produce significant variation in potency within and across target sets have also been investigated [19–22]. In addition, MMP-based analyses have also been performed to assess the effects of replacing various chemical groups on different experimentally determined and calculated properties [23–26]. MMP-based analyses are devoid of the "black box" nature often associated with other computational approaches. In these cases, the association between biological activity and molecular structure is evident and interpretable in an intuitive manner [13].

## Methods for Dimension Reduction

Projection of compounds into a chemical reference space represented either by molecular fingerprints or descriptors is often a prerequisite for computational analyses. However, such reference spaces are high-dimensional and as such

their intuitive depiction is rather difficult. Reduction of these multi-dimensional spaces to two or three dimensions is often performed in order to ease their navigation. The resulting low-dimensional data is used to represent bioactive compounds which can then be readily visualized by routinely used methodologies. However, molecular structures need to be examined separately and extraction of pertinent SAR information requires chemical expertise [27]. Transformation of multivariate data into a space of lower dimensionality is frequently referred to as nonlinear mapping and represents one possible approach to dimension reduction [28]. The primary objective of nonlinear mapping is the conservation of neighborhood relationships such that proximity in multi-dimensional space is reproduced in the lower dimensions [29]. To this end, several mathematical techniques have been applied to perform dimension reduction [28, 30].

Another popular dimension reduction technique is principal component analysis (PCA), which generates linear orthogonal combinations of original descriptor sets. The smaller set of novel variables generated by PCA is sufficient to account for a certain degree of variance produced by the original descriptor set [27]. PCA results in a coordinate-based low-dimensional reference space and can be applied to large data sets. By contrast, multi-dimensional scaling (MDS), a classical example of the nonlinear mapping technique, is used for the transformation of the *coordinate-free reference space* obtained by pairwise molecular fingerprint comparisons. MDS is better suited for preserving similarity relationships while decreasing dimensionality, although it is less favorable for large compound sets due to computational challenges. This issue can be circumvented by using MDS in combination with feed-forward neural networks [30]. Alternate approaches to dimension reduction also include Kohonen networks or self-organizing maps (SOMs) [31]. Irrespective of the dimension reduction technique used to transform multi-dimensional data, one can only minimize but never completely avoid the associated loss of information.

## Attributes of Structure-Activity Relationships

SAR characteristics of bioactive compound sets are determined by the degree of change in activity accompanied by their structural modifications [32]. When clear trends in bioactivity arise due to systematic chemical changes of bioactive

compounds, they represent "continuous" SARs [33]. The presence of structurally similar compounds with comparable potencies is indicative of continuous SARs. Therefore, such SARs are consistent with the SPP and constitute a *global* molecular similarity perspective [32]. Additionally, structural modifications may also lead to increasingly diverse compounds with conserved activity, a phenomenon known as *scaffold hopping* [34]. In such cases, these compounds often have similar shapes or pharmacophores that represent *local* activity-relevant similarities. Thus, scaffold hopping also falls within the spectrum of continuous SARs. Conversely, if minor chemical replacements induce large changes in activity within a compound set, the underlying SAR is said to be "discontinuous" [33]. The distinguishing feature of discontinuous SARs is the presence of structurally similar compounds with significantly different potencies. Such pairs of compounds are often referred to as *activity cliffs* [35]. Discontinuous SARs fall outside the SPP applicability domain and often pose an impediment to molecular similarity analysis. However, these two SAR categories do not necessarily occur independently of each other. Rather, continuous and discontinuous SAR elements often co-exist within compound sets and consequently, the ensuing SAR category is "heterogeneous" [36]. In general, the global SAR for a set of compounds that share activity against a given target, i.e. an activity class, can belong only to one of the above mentioned categories [33].

## Conventional SAR Analysis

In medicinal chemistry, SARs are typically investigated on a *case-by-case* basis and the analysis entails studying structurally similar compound series active against a biological target. Exploration of closely related series is carried out to understand how structural perturbations influence the bioactivity of compounds. Such investigations usually involve manual comparison of the 2D molecular graphs of bioactive analogs. The analogs are often represented in a tabular format as core structures (or scaffolds) and various substituents, along with their biological activities. Such R-group tables are intuitive tools most commonly used in SAR analysis. These are also suitable to determine SAR trends that aid in compound design and lead optimization [37].

Despite their clear merits, these R-group tables become increasingly difficult to interpret as the number of analogs increases. Moreover, such traditional SAR analyses rely heavily on the experience and intuition of medicinal chemists. As a result, the outcome is often subjective and prone to inconsistencies [38]. Generation of core and R-group matrices using a computational approach has also been performed [39]. Numerous other representations, like tree maps and radial clustergrams, that depict structural similarity and bioactivity distribution as well as other molecular properties have also been designed [40, 41]. Recently, MMP-based SAR matrices that capture SAR information content in large compound sets in various intuitive ways have been reported [42]. Although, computational methods can be utilized to organize large compound sets into SAR tables, the chemical structures of individual molecules may still require a thorough examination.

In order to facilitate derivation of quantitative SAR information, mathematical functions are employed that relate structural features and properties of compounds to their activity. Such methodologies follow the quantitative SAR (QSAR) analysis paradigm [43, 44]. Despite variations in their design, the primary objective of QSAR methods is to facilitate activity prediction for novel compounds. QSAR models were initially generated using linear 2D approaches, but nonlinear as well as 3D modeling have also been attempted [44–48]. Recent advances incorporating machine-learning techniques and artificial intelligence have resulted in QSAR methodologies with improved prediction capabilities [48, 49].

An inherent limitation common to all QSAR methodologies is that their application is confined to congeneric compound series, i.e. compounds that bear close structural resemblance. Thus, other compounds with dissimilar structures fall outside the applicability domain of QSAR models [50]. Even within the applicability domain, credibility of QSAR modeling is only ensured when the underlying SAR of the compounds is continuous. Presence of activity cliffs often impede the success of QSAR for which predictions can be inconsistent [35]. Nevertheless, activity cliffs are considered important by medicinal chemists as they serve as centers on which hit-to-lead and lead optimizations studies can be focused in order to obtain compounds with improved bioactivity [2, 35].

## Activity Landscape Representation

SARs for different compound sets are often described using the activity landscape concept modeled after actual geographic landscapes [51]. An activity landscape representation combines chemical similarities and activity differences between compounds active against a given biological target. Compounds that constitute the chemical reference space are positioned along the xy-plane in a way that captures molecular similarity. Thus, structurally related compounds are proximal in this two-dimensional projection while dissimilar compounds are separated from each other. Activity information pertaining to every constituent compound is incorporated as the third dimension.

The result can be envisioned as a topological surface with variable levels of elevation [52]. Accordingly, structural alterations of compounds would constitute transitions in the chemical space and the resulting effects on activity may be perceived as variations in surface elevation. Therefore, small chemical transformations accompanied by small potency changes, i.e. SAR continuity, would produce a smooth activity landscape. Alternatively, SAR discontinuity, which is typified by minor structural modifications leading to large potency differences, would generate a rugged landscape [35, 52]. An activity landscape containing smooth regions interspersed with rugged topological features would represent a heterogeneous SAR [32, 33].

These idealized activity landscapes are shown in **Figure 1**. It is, however, important to note that representation of SARs as well as the rationalization of their information content is far from trivial. For medicinal chemists, SARs with predictable potency progression are of high interest in compound design. In such cases, SAR continuity is an essential consideration. Moreover, continuous SARs are also relevant when multiple starting points are required for hit-to-lead studies. However, when the focus shifts to lead optimization, SAR discontinuity is also important and activity cliffs are considered. Thus, methodologies that link SAR continuity and discontinuity are an implicit requirement for SAR exploration and exploitation [37, 38]. Such approaches are often referred to as *SAR profiling* methods.

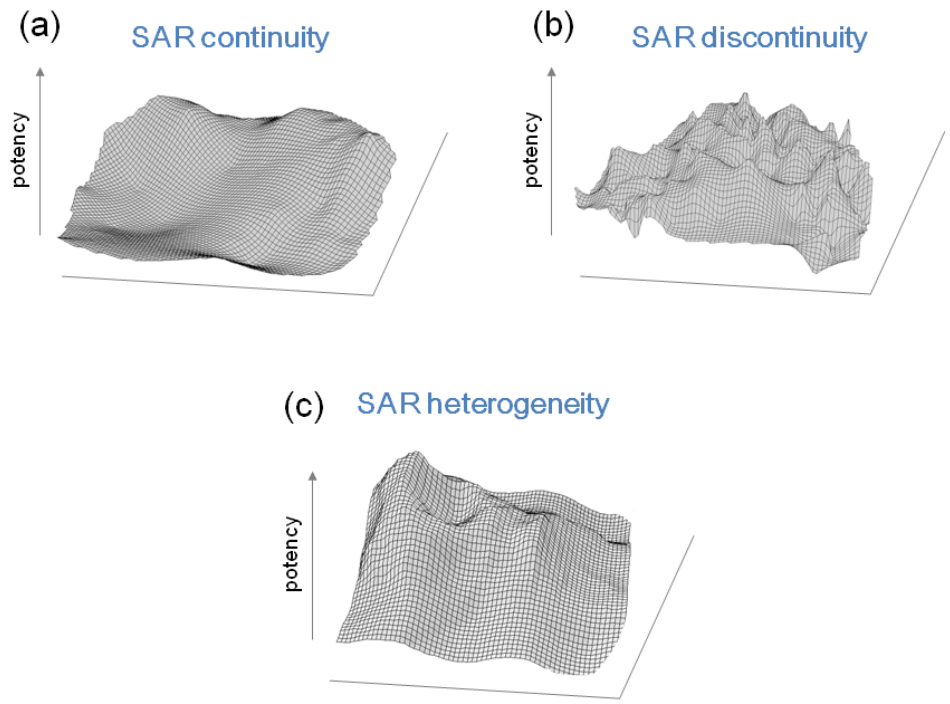



**Figure 1: Idealized activity landscapes.** Hypothetical activity landscape representations exhibiting (a) SAR continuity, (b) discontinuity and (c) heterogeneity are shown. These hypersurfaces are generated by projecting compounds into xy-plane derived from chemical reference space, followed by the addition of potency data as the z-axis. Here, increase in the distances along the 2D plane reflect decrease in chemical similarity and potency distribution is related to surface elevation. (*adapted from Wassermann et al.*[38])

## Numerical Functions for SAR Analysis

Numerical SAR analysis functions like the SAR index (SARI) and structure-activity landscape index (SALI) quantify SAR features by taking into account pairwise structural similarities and potency differences within compound sets [53, 54]. By systematic evaluation of structural similarity and activity distribution within data sets, these functions provide direct access to various SAR relevant characteristics. SARI comprises of two separately calculated components, the continuity and the discontinuity scores. Raw continuity scores are

derived from potency weighted average of pairwise chemical similarities and calculated as

$$raw_{\text{cont}} = 1 - \frac{\sum_{\text{compounds i}\neq\text{j}} w_{\text{ij}} \times sim(i,j)}{\sum_{\text{compounds i}\neq\text{j}} w_{\text{ij}}}$$

where the weight $w_{\text{ij}}$ for each compound pair $(i,j)$ is set to

$$w_{\text{ij}} = \frac{pot(i) \times pot(j)}{1 + \mid pot(i) - pot(j) \mid}$$

Raw discontinuity scores are generated as follows

$$raw_{\text{disc}} = \frac{\sum_{\{\text{i,j}|\text{sim(i,j)}>\text{thres, i}\neq\text{j}\}} \mid pot(i) - pot(j) \mid \times sim(i,j)}{\mid \{i,j \mid sim(i,j) > thres,\ i \neq j\} \mid}$$

Here, $pot(i)$, $pot(j)$ represent the potency values of compounds $i$ and $j$, $sim(i,j)$ denotes their similarity value and $thres$ corresponds to a predefined similarity threshold. The raw scores are transformed to the value range [0,1] after statistical normalization. SARI is calculated as the mean between the continuity score and the complement of the discontinuity score

$$\text{SARI} = \frac{1}{2}(cont_{\text{norm}} + 1 - disc_{\text{norm}})$$

where $cont_{\text{norm}}$ and $disc_{\text{norm}}$ are the normalized continuity and discontinuity scores. Therefore, high, intermediate and low SARI scores are indicative of global SAR continuity, heterogeneity and discontinuity, respectively.

The objective of SALI scoring function is to prioritize potency differences of large magnitude between structurally similar compounds and the scores are calculated as

$$\text{SALI(i,j)} = \frac{pot(i) - pot(j)}{1 - sim(i,j)}$$

SALI scores are designed to describe activity cliffs of varying magnitude in compound data sets. Although, both SALI and SARI discontinuity scores encode activity cliff information, unlike SARI discontinuity scores that can have a maximum value of unity, SALI scores may have a value range of $[0,\infty]$. Moreover, SALI scores are local in nature while SARI scores are global [27].

## SAR Visualization Techniques

Numerous attempts have been made in the SAR visualization area to systematically identify relevant features in large sets of compounds with activity annotations. Such tools also allow intuitive and interpretable representation of SARs [27]. For example, structure-activity similarity (SAS) maps constitute a 2D graphical representation where pairwise structural and activity similarities between compounds are plotted along x- and y-axes respectively [51]. A variant of SAS maps that accounts for molecular properties has also been designed [55].

Molecular network representations such as network-like similarity graphs (NSGs) also constitute a popular SAR visualization technique [56, 57]. Like SAS maps, NSGs are graphical networks in which compounds are depicted as nodes and similarity relationships between them as edges. Edges are drawn only if pairwise similarities exceed a predefined threshold. Per-compound discontinuity score calculated as

$$raw_{\text{disc}}(i) = \frac{\sum_{\{\text{i,j}|\text{sim(i,j)}>\text{thres, i}\neq\text{j}\}} \Delta pot(i,j) \times sim(i,j)}{\mid \{i,j \mid sim(i,j) > thres,\ i \neq j\} \mid}$$

determines the node size where $sim(i,j)$, $\Delta pot(i,j)$ denote the chemical similarity and potency difference between compounds $i$ and $j$ while $thres$ corresponds to the similarity threshold. Potency data is encoded as the node color. Additionally, compound clustering is performed and cluster SARI discontinuity scores calculated to identify individual groups with high SAR discontinuity. NSGs have also been successfully used to automatically extract pertinent SAR information from high-throughput screening data [58]. An exemplary NSG and the various elements of its design are reported in **Figure 2**. These network-based landscape models are designed to study both global as well as local SAR characteristics [56]. Other network representations like similarity-potency trees (SPTs) are centered on individual compounds and provide a local view of SARs [59]. SPTs are generated for individual compounds in a data set and ranked according to their local SAR information content. Such a systematic exploration of SPTs limits the loss of SAR information in data analysis [38]. Similar analyses of SARs in the vicinity of reference compounds can also be carried out with

the help of chemical neighborhood graphs (CNGs) [60]. CNGs are very useful for analyzing complex SAR features and provide multiple local SAR views [27].

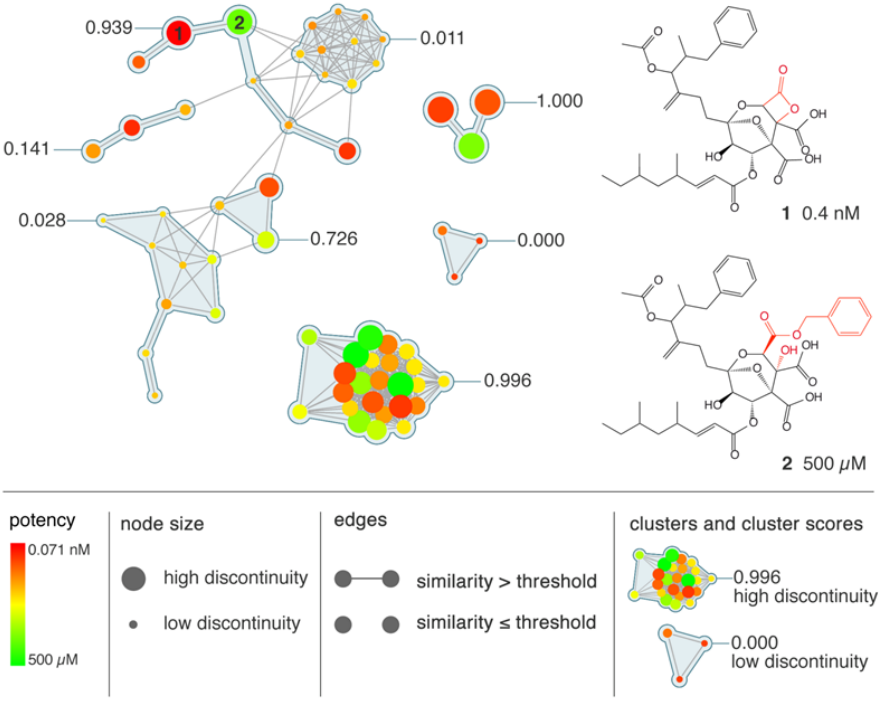



**Figure 2: Single-target activity landscape representation.** An exemplary NSG for a set of 71 squalene synthase inhibitors is shown. The principal design elements are described in the legend below the graph. Compound subsets identified by hierarchical clustering are displayed against a light blue background and annotated with cluster discontinuity scores. A compound pair forming an activity cliff is highlighted in the graph (labeled 1 and 2) and their structures as well as potencies are reported. (*taken from Wassermann et al.*[38])

Most SAR visualization tools are designed to enable the analysis of large sets of compounds. However, lead optimization approaches usually require the exploration of analog series. For this purpose, combinatorial analog graphs (CAGs) have been introduced. CAGs systematically organize analog series according to substitution site combinations on the basis of R-group decomposition [61]. Substitution patterns that produce SAR discontinuity and possible yet unexplored analogs can be easily identified.

Graphical SAR representations based on calculated structural similarities often require close inspection of compound structures to rationalize the SAR information content. This inherent limitation can be circumvented by utilizing

well-defined substructure relationships instead of calculated similarities. Such substructure relationships can be systematically generated for compounds comprising a data set using MMPs. Compounds that differ by a single substructure, are further organized into matching molecular series (MMS). These MMS are represented in a network representation known as the bipartite matching molecular series graph (BMMSG) [62].

Substructure-based approaches focus on compound design strategies that associate structural fragments with bioactivity information. Substructures can either be predefined or generated systematically from compounds sets by first removing all side chains, followed by iterative pruning of rings. The latter approach results in the generation of molecular frameworks or *scaffolds* that can be annotated with activity information of the compounds from which they were obtained and organized into a hierarchy [63, 64]. Chemical space traversal using such scaffold hierarchies can aid in compound design [65].

## Multi-Target SAR Analysis

SAR investigations routinely focus on sets of compounds that are active against specific targets with the objective of yielding novel compounds with improved potency [66]. Many compound sets are also active against more than one target, thereby, forming multi-target SARs and techniques that aid in their analyses need to be developed.

Adaptation of the activity landscape concept to systematically account for dual target activities of compounds in the form of potency ratios has recently been attempted using NSGs [67]. Thus, the resultant NSGs form a selectivity landscape. **Figure 3** illustrates the design as well as rationalization of selectivity NSGs and indicates the conceptual difference with respect to NSGs generated for single targets. SAS maps have also been extended to accommodate compound selectivity information [68]. Compound selectivity analysis has also been carried out in analog series such that R-groups are expressed as predefined pharmacophore features and similarity is assessed locally in the form of pharmacophore edit distances [69]. Pairs of structurally similar compounds with a large difference in their target selectivity, referred to as selectivity cliffs, form the most prominent features of such landscapes.

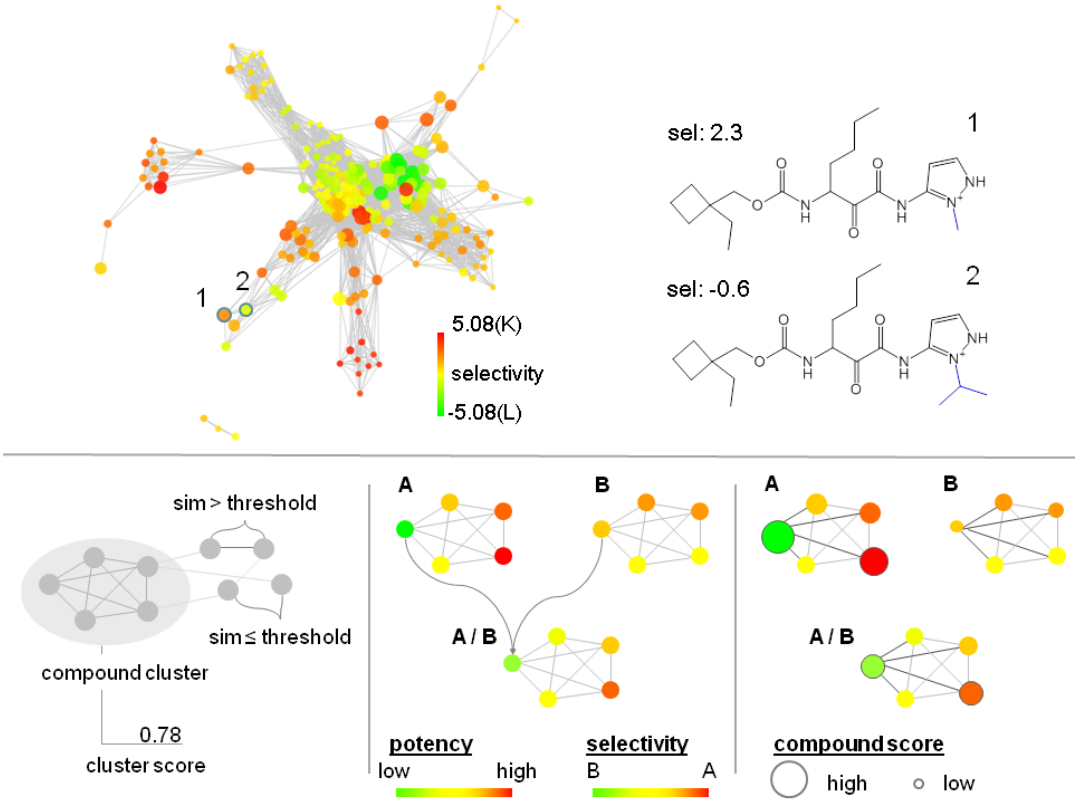



**Figure 3: From activity to selectivity landscapes.** An exemplary selectivity NSG for a set of 234 inhibitors active against cathepsins K and L is shown. The principal design elements are described in the legend below the graph. A selectivity cliff formed by a compound pair (labeled 1 and 2) is highlighted and the structures are shown. In addition, selectivities (i.e. potency ratios) are reported. (*adapted from Wassermann et al.*[38] *and Peltason et al.*[67])

Efforts to generate graphical activity landscape representations for compound sets with activity against three or more targets have also been made. Similarity relationships for such multi-target sets are depicted using a regular NSG and a potency binning scheme is used to generate compound activity profiles that are then provided as node annotations [70]. Multi-target discontinuity scores that quantitatively compare the potency differences of compounds with their structural neighbors across multiple targets in a pairwise manner are used to scale the nodes. The elements of multi-target NSG generation as well as an example is shown in **Figure 4**.

Compound activity profile encoding also facilitates the identification of single and multi-target cliffs and has been employed to systematically analyze such cliffs in publicly available bioactive compounds [71].

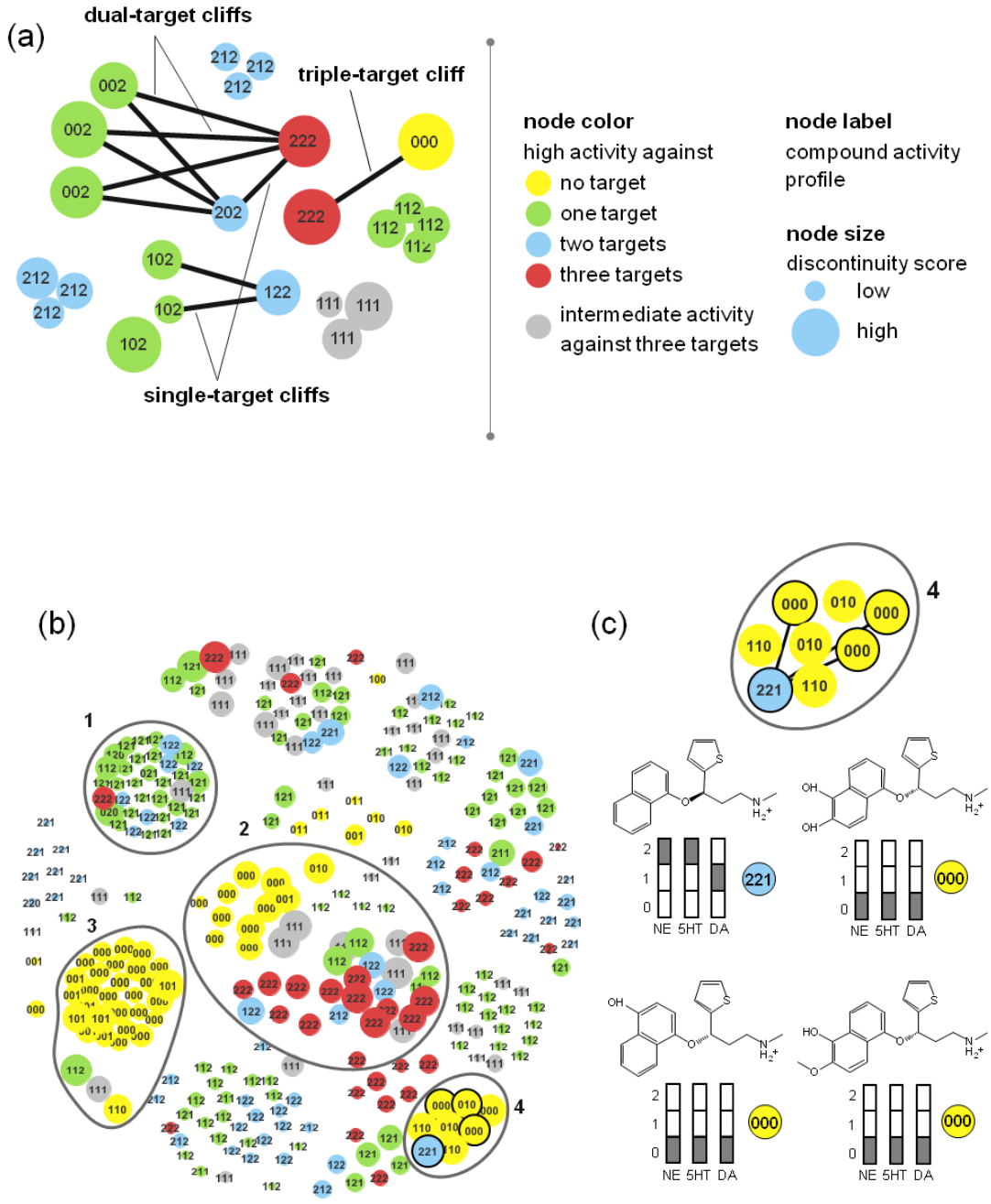



**Figure 4: Multi-target activity landscape.** Figure 4 (a) explains the details of various features present in multi-target network-based landscape design. An exemplary multi-target NSG for a set of 299 monoamine transporter inhibitors is displayed in (b). Selected compound subsets with multi-target cliffs are encircled and numbered. An enlarged view of cluster 4 containing a dual-target cliff is shown in (c). Structures and activity profiles of compounds representing the dual-target cliff are also reported. (*adapted from Dimova et al.*[70])

A second numerical function to assess the ability of compounds to distinguish between various targets within target families has recently been reported [72]. In addition, SAS maps have been modified to incorporate multi-target activity information by calculating activity similarity between vectors of compound potencies against multiple targets [73].

Multi-target activity landscapes designed so far have an inherent limitation that they become increasingly difficult to interpret with increasing numbers of targets. Moreover, calculation of activity similarity potentially also results in loss of SAR information. Second-generation multi-target activity landscape models have been introduced in order to circumvent such limitations [74].

This 3D multi-target activity landscape combines chemical and target spaces in circular representations supporting interactive analysis of projected compounds. Compounds with clearly defined selectivity patterns and structure-activity profiles can be identified. However, multi-target graphical representations require that compounds comprising the data sets have potency annotations for all the targets under consideration. Thus, they are not capable of handling incomplete activity matrices. In addition to various multi-target graphical representations, various systematic analyses at the level of molecular scaffolds have also been performed to account for multi-target activity information. Such studies have led to the identification of scaffolds selective for closely related targets [75] as well as those that are promiscuously active across multiple target families [76].

## Thesis Outline

The primary objective of this dissertation is the development of methodologies for systematic single and multi-target SAR analyses. The dissertation consists of eight individual chapters that form a sequence of studies.

*Chapter 1* of this dissertation reports the design of 3D activity landscapes for compound data sets. *Chapter 2* provides a comparison of 3D activity landscapes with 2D landscape representations (NSGs). *Chapter 3* reports the application of conditional feature probability calculations for individual compounds in ligand data sets to provide a higher resolution graphical analysis of SAR relevant characteristics.

*Chapters 4* and *5* introduce graphical methodologies to analyze compounds with different mechanisms of action for a target receptor and identify structural changes that lead to mechanistic changes.

A novel multi-target activity landscape representation generated using SOMs that encodes target selectivity profiles of compounds is presented in *Chapter 6*. Furthermore, the development of a second multi-target activity landscape that is suitable for data sets with incomplete multi-target activity annotations is introduced in *Chapter 7*. Assessment of differentiation potential of imidazole-based inhibitors for various kinases is reported in *Chapter 8*.

# Bibliography

[1] Johnson M., Maggiora G. M. *Concepts and applications of molecular similarity.*, John Wiley and Sons, New York, USA, **1990**.

[2] Kubinyi H. Similarity and dissimilarity: a medicinal chemist's view. *Perspect. Drug Discov. Des.*, **1998**, 9-11, 225-252.

[3] Martin Y. C., Kofron J. L., Traphagen L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.*, **2002**, 45, 4350-4358.

[4] Peltason L., Bajorath J. Molecular similarity analysis in virtual screening. In Varnek A. and Tropsha A. (Eds.), *Chemoinformatics: An Approach to Virtual Screening*, RSC Publishing, Cambridge, UK, **2008**, 120-149.

[5] Bajorath J. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comput. Sci.*, **2001**, 41, 233-245.

[6] Willett P. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.*, **1998**, 38 (6), 983-996.

[7] Xue L., Bajorath J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Combin. Chem. High Throughput Screen.*, **2000**, 3, 363-372.

[8] MACCS Structural keys. Symyx Software, San Ramon, CA, USA.

[9] Bender A., Mussa H. Y., Glen R. C., Reiling S. Molecular similarity searching using atom environments, information-based feature selection, and a naive bayesian classifier. *J. Chem. Inf. Comput. Sci.*, **2004**, 44, 170-178.

[10] Rogers D., Hahn M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, **2010**, 50, 742-754.

[11] McGregor M. J., Muskal S. M. Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. *J. Chem. Inf. Model.*, **1999**, 39 (3), 569-574.

[12] Maggiora G. M., Shanmugasundaram V. Molecular similarity measures. In Bajorath J. (Ed.), *Methods in Molecular Biology, vol 275: Chemoinformatics: Concepts, Methods and Tools for Drug Discovery*, Humana Press, Totowa, New Jersey, USA, **2004**, 1-50.

[13] Wassermann A. M., Dimova, D., Iyer P. and Bajorath J. Advances in computational medicinal chemistry: matched molecular pair analysis. *Drug Develop. Res.*, **2012**, 73, 518-527.

[14] Papadatos G., Alkarouri M., Gillet V. J., Willett P., Kadirkamanathan V., Luscombe C. N., Bravi G., Richmond N. J., Pickett S. D., Hussain J., Pritchard J. M., Cooper A. W., Macdonald S. J. Lead optimization using matched molecular pairs: inclusion of contextual information for enhanced prediction of hERG inhibition, solubility, and lipophilicity. *J. Chem. Inf. Model.*, **2010**, 50, 1872-1876.

[15] 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.

[16] Sheridan R. P. The most common chemical replacements in drug-like compounds. *J. Chem. Inf. Comput. Sci.*, **2002**, 42, 103-108.

[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] Wassermann A. M., Bajorath J. Identification of target family directed bioisosteric replacements. *Med. Chem. Commun.*, **2011**, 2, 601-606.

[19] 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.

[20] Stumpfe D., Bajorath J. Exploring activity cliffs in medicinal chemistry. *J. Med. Chem.*, **2012**, 55, 2932-2942.

[21] 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.

[22] Hu Y., Bajorath J. Chemical transformations that yield compounds with distinct activity profiles. *ACS Med. Chem. Lett.*, **2011**, 2, 523-527.

[23] Leach A. G., Jones H. D., Cosgrove D. A., Kenny P. W., Ruston L, MacFaul P, Wood J. M., Colclough N., Law B. Matched molecular pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and oral exposure. *J. Med. Chem.*, **2006**, 46, 6672-6682.

[24] Gleeson P., Bravi G., Modi S., Lowe D. ADMET rules of thumb II: a comparison of the effects of common substituents on a range of ADMET parameters. *Bioorg. Med. Chem. Lett.*, **2009**, 17, 5906-5919.

[25] Lewis M. L., Cuchurall-Sanchez L. Structural pairwise comparisons of HLM stability of phenyl derivatives: introduction of Pfizer metabolism index (PMI) and metabolism-lipophilicity efficiency (MLE). *J. Comput. Aided. Mol. Des.*, **2009**, 23, 97-103.

[26] Schultes S., de Graaf C., Berger H., Mayer M., Steffen A., Haaksma E. E. J., de Esch I. J. P., Leurs R., Krämer O. A medicinal chemistry perspective on melting point: matched molecular pair analysis of the effects of simple descriptors on the melting point of drug-like compounds. *Med. Chem. Commun.*, **2012**, 3, 584-591.

[27] 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 Discov. Today*, **2010**, 15, 630-639.

[28] Hair J. F., Anderson R. H., Tatham R. L., Black W. C. *Multivariate data analysis.*, Prentice Hall, New Jersey, USA, **1998**.

[29] Gedeck P., Willett P. Visual and computational analysis of structure-activity relationships in high-throughput screening data. *Curr. Opin. Chem. Biol.*, **2001**, 5, 389-395.

[30] Agrafiotis D. K., Lobanov V. S. Nonlinear mapping networks. *J. Chem. Inf. Comput. Sci.*, **2000**, 40 (6), 1356-1362.

[31] Kohonen T. *Self-organizing maps*, Springer, Heidelberg, Germany, **1996**.

[32] Eckert H., Bajorath J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov. Today.*, **2007**, 12 (5-6), 225-233.

[33] Peltason L., Bajorath J. Systematic computational analysis of structure-activity relationships: concepts, challenges, and recent advances. *Future Med. Chem.*, **2009**, 1 (3), 451-466.

[34] Böhm H. -J., Flohr A., Stahl M. Scaffold hopping. *Drug Discov. Today: Technol.*, **2004**, 1, 217-224.

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

[36] Peltason L., Bajorath J. Molecular similarity analysis uncovers heterogeneous structure-activity relationships and variable activity landscapes. *Chem. Biol.*, **2007**, 14, 489-497.

[37] Stumpfe D., Bajorath J. Methods for SAR visualization. *RSC Adv.*, **2012**, 2, 369-378.

[38] Wassermann A. M., Waver M., Bajorath J. Activity landscape representations for structure-activity relationship analysis. *J. Med. Chem.*, **2010**, 53, 8209-8223.

[39] Agrafiotis D. K., Shemanarev M., Connolly P. J., Farnum M., Lobanov, V. S. SAR maps: a new SAR visualization technique for medicinal chemists. *J. Med. Chem.*, **2007**, 50, 5926-2937.

[40] Kibbey C., Calvet A. Molecular Property eXloprer: a novel approach to visualizing SAR using tree-maps and heatmaps. *J. Chem. Inf. Model.*, **2005**, 45 (2), 523-532.

[41] Agrafiotis D. K., Bandyopadhyay D., Farnum M., Radial clustergrams: visualizing the aggregate properties of hierarchical clusters. *J. Chem. Inf. Model.*, **2007**, 47, 69-75.

[42] Wassermann A. M., Haebel P., Weskamp N., Bajorath J. SAR matrices: automated extraction of information-rich SAR tables from large compound data sets. *J. Chem. Inf. Model.*, **2012**, 52 (7), 1769-1776.

[43] van de Waterbeemd H., Rose S. Quantitative approaches to structure-activity relationships. In Wermuth C. G. (Ed.), *The Practice of Medicinal Chemistry, 3rd ed.*, Academic Press, Burlington, MA, USA, **2008**, 491-513.

[44] Esposito E. X., Hopfinger A. J., Madura J. D. Methods for applying the quantitative structure-activity relationship paradigm. *Methods Mol. Biol.*, **2004**, 275, 131-213.

[45] Kubinyi H. Quantitative structure-activity relationships. 7. The bilinear model, a new model for nonlinear dependence of biological activity on hydrophobic character. *J. Med. Chem.*, **1977**, 20, 625-629.

[46] Manallack D. T., Eliis D. D., Livingstone D. J. Analysis of linear and nonlinear QSAR data using neural networks. *J. Med. Chem.*, **1994**, 37, 3758-3767.

[47] Kubinyi H. QSAR and 3D QSAR in drug design. Part 1: Methodology. *Drug Discov. Today*, **1997**, 2, 457-467.

[48] Michielan L., Moro S. Pharmaceutical perspectives of nonlinear QSAR strategies. *J. Chem. Inf. Model.*, **2010**, 50, 961-978.

[49] Geppert H., Vogt M., Bajorath J. Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inf. Model.*, **2010**, 50, 205-216.

[50] Dimitrov S., Dimitrova G., Pavlov T., Dimitrova N., Patlewicz G., Niemala J., Mekenyan O. A stepwise approach for defining the applicability domain of SAR and QSAR models. *J. Chem. Inf. Model.*, **2005**, 45, 839-849.

[51] Shanmugasundaram V., Maggiora G. M. Characterizing property and activity landscapes using an information-theoretic approach. *222nd ACS National Meeting.*, **2001**, Division of Chemical Information, Abstract no. 77.

[52] Bajorath J., Peltason L., Wawer M., Guha R., Lajiness M. S., Van Dire J. H. Navigating structure-activity landscapes. *Drug Discov. Today.*, **2009**, 14 (13-14), 698-705.

[53] Peltason L., Bajorath J. SAR index: quantifying the nature of structure-activity relationships. *J. Med. Chem.*, **2007**, 50, 5571-5578.

[54] Guha R., Van Drie J. H. Structure-activity landscape index: identifying and quantifying activity cliffs. *J. Chem. Inf. Model*, **2008**, 48, 646-658.

[55] 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.

[56] 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.

[57] Wawer M., Peltason L., Bajorath J. Elucidation of structure-activity relationship pathways in biological screening data. *J. Med. Chem.*, **2009**, 52, 1075-1080.

[58] Wawer M., Bajorath J. Extraction of structure-activity relationship information from high-throughput screening data. *Curr. Med. Chem.*, **2009**, 16, 4049-4057.

[59] Wawer M., Bajorath J. Similarity-Potency Trees: a method to search for SAR information in compound data sets and derive SAR rules. *J. Chem. Inf. Model.*, **2010**, 50, 1395-1409.

[60] Wawer M., Sun S., Bajorath J. Computational characterization of SAR microenvironments in high-throughput screening data. *Intl. J. High Throughput Screen*, **2010**, 1, 15-27.

[61] Peltason L., Weskamp N., Teckentrup A., Bajorath, J. Exploration of structure-activity relationship determinants in analogue series. *J. Med. Chem.*, **2009**, 52, 3212-3224.

[62] Wawer M., Bajorath J. Local structural changes, global data views: graphical substructure-activity relationship trailing. *J. Med. Chem.*, **2011**, 54, 2944-2951.

[63] Schffenhauer A., Ertl P., Roggo S., Wetzel S., Koch M. A., Waldmann H. The scaffold tree - visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inf. Model.*, **2007**, 47, 47-58.

[64] Agrafiotis D. K., Wiener J. J. M. Scaffold Explorer: an interactive tool for organizing and mining structure-activity data spanning multiple chemotypes. *J. Med. Chem.*, **2010**, 53 (13), 5002-5011.

[65] Renner S., van Otterlo W. A. L., Dominguez Seoane M., Möcklinghoff S., Hofman B., Wetzel S., Schffenhauer A., Ertl. P., Oprea T. I., Steinhilber D., Brunsveld L., Rauh D., Waldmann H. Bioactivity-guided mapping and navigation of chemical space. *Nat. Chem. Biol.*, **2009**, 5, 585-592.

[66] Hopkins A. L. Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.*, **2008**, 4, 682-690.

[67] Peltason L., Hu Y., Bajorath J. From structure-activity to structure-selectivity relationships: quantitative assessment, selectivity cliffs, and key compounds. *Chem. Med. Chem.*, **2009**, 4, 1864-873.

[68] Perez-Villanueva J., Santos R., Hernandez-Campos A., Giulianotti M. A., Castillo R., Medina-Franco J. L. Structure-activity relationships of benzimidazole derivatives as antiparasitic agents: Dual-activity difference (DAD) maps. *Med. Chem. Commun.*, **2011**, 2, 44-49.

[69] 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. *Chem. Med. Chem.*, **2010**, 5, 847-858.

[70] Dimova D., Wawer M., Wassermann A. M., Bajorath J. Design of multitarget activity landscapes that capture hierarchical activity cliff distributions. *J. Chem. Inf. Model.*, **2011**, 51 (2), 258-266.

[71] 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.

[72] Dimova D., Bajorath J. Computational chemical biology: identification of small molecular probes that discriminate between members of target families. *Chem. Biol. Drug. Des.*, **2012**, 79, 369-375.

[73] Waddell J., Medina-Franco J. L. Bioactivity landscape modeling: chemoinformatic characterization of structure-activity relationships of compounds tested across multiple targets. *Bioorg. Med. Chem.*, **2012**, 20, 5443-5452.

[74] de la Vega de León A., Bajorath J. Design of a three-dimensional multitarget activity landscape. *J. Chem. Inf. Model.*, **2012**, 52 (11), 2876-2883.

[75] Hu Y., Wassermann A. M., Lounkine E., Bajorath J. Systematic analysis of public domain compound potency data identifies selective molecular scaffolds across druggable target families. *J. Med. Chem.*, **2010**, 53, 752-758.

[76] 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.

# Chapter 1

# Rationalizing three-dimensional activity landscapes and the influence of molecular representations on landscape topology and the formation of activity cliffs

## Introduction

Systematic structural similarity and activity relationships can be captured and represented using the activity landscape concept. Activity landscapes are analogous to geographical maps and intuitively characterize the SARs underlying sets of bioactive molecules. Various attributes of SAR have previously been elucidated using theoretical 3D models. Nevertheless, such models have not been generated for actual compound sets. In the following, generation of real 3D activity landscapes using a novel computational approach is reported. The methodology has been applied to various activity-annotated compound sets including a high-throughput screening data set. In addition, three conceptually different molecular representations have been used for landscape generation.

# Rationalizing Three-Dimensional Activity Landscapes and the Influence of Molecular Representations on Landscape Topology and the Formation of Activity Cliffs

Lisa Peltason,[†] Preeti Iyer,[†] 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

Received March 5, 2010

Activity landscapes are defined by potency and similarity distributions of active compounds and reflect the nature of structure−activity relationships (SARs). Three-dimensional (3D) activity landscapes are reminiscent of topographical maps and particularly intuitive representations of compound similarity and potency distributions. From their topologies, SAR characteristics can be deduced. Accordingly, idealized theoretical landscape models have been utilized to rationalize SAR features, but "true" 3D activity landscapes have not yet been described in detail. Herein we present a computational approach to derive approximate 3D activity landscapes for actual compound data sets and to analyze exemplary landscape representations. These activity landscapes are generated within a consistent reference frame so that they can be compared across different activity classes. We show that SAR features of compound data sets can be derived from the topology of landscape models. A notable correlation is observed between global SAR phenotypes, assigned on the basis of SAR discontinuity scoring, and characteristic landscape topologies. We also show that different molecular representations can substantially alter the topology of activity landscapes for a given data set and modulate the formation of activity cliffs, which represent the most prominent landscape features. Depending on the choice of molecular representations, compounds forming a steep activity cliff in a given landscape might be separated in another and no longer form a cliff. However, comparison of alternative activity landscapes makes it possible to focus on compound subsets having high SAR information content.

## INTRODUCTION

The concept of activity landscapes plays a key role in understanding structure−activity relationships (SARs).[1−3] Activity landscapes are best rationalized as hypersurfaces in biologically relevant chemical space, where biological activity (compound potency) adds another dimension.[3] The interpretation of high-dimensional activity landscapes is generally difficult and, consequently, two- and three-dimensional (2D and 3D, respectively) representations of activity landscapes have been taken into consideration. If we envision a 2D projection of chemical space with compound potency added as a third dimension, then activity landscapes become reminiscent of geographical maps that can readily be interpreted.[2,3] Smooth regions that are reminiscent of rolling hills[1] correspond to areas where gradual changes in chemical structure are accompanied by moderate changes in biological activity. Compounds mapping to such areas are related by so-called continuous SARs.[3] By contrast, rugged regions in activity landscapes that are canyon-like[1] correspond to areas where small chemical changes have dramatic effects on the biological response, and hence, compounds mapping to these areas form discontinuous SARs.[3] The strongest articulation of SAR discontinuity are so-called activity cliffs[1] that are formed by pairs of structurally very similar compounds with large differences in potency, i.e., small steps in chemical space are accompanied by large changes in activity.

Numerical analysis functions including the SAR index (SARI)[4] or the structure−activity landscape index (SALI)[5] have been introduced to characterize global SAR features present in compound data sets on a large scale[4] and to quantify SAR discontinuity.[4,5] These analysis functions systematically relate compound similarity and potency to each other and can also be applied to quantify how well a computational model fits a given activity landscape.[6] In combination with similarity-based molecular network representations,[5,7] these calculations make it possible to identify and compare activity cliffs in compound data sets. Annotating or combining network representations, such as SALI maps[5] or network-like similarity graphs[7] (NSGs), with potency and SAR continuity and/or discontinuity score[4,5] information enables the 2D representation of activity landscapes, including the identification of compounds that are related by continuous or discontinuous SARs, and the comparison of global and local SAR features. Systematic NSG analysis has revealed that a significant degree of SAR heterogeneity exists in most compound data sets, due to the presence of different continuous and discontinuous local SARs.[7,8] Activity cliffs of varying magnitude can essentially be found in compound data sets of any source, including raw screening data, irrespective of the nature of the biological targets.[7−9] It follows that most activity landscapes are likely to display variable topology, i.e., in terms of an idealized 3D landscape model, they consist of smooth rolling hill-type regions that are interspersed with cliff areas and canyons. Such variable activity landscapes provide the basis for the identification of structurally diverse compounds having similar activity (in

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

[†] These authors contributed equally to this paper.

10.1021/ci100091e 
Published on Web 05/05/2010

smooth regions) and for the optimization of compound potency (at activity cliffs).[3]

It is also well-appreciated that the nature of activity landscapes is much influenced by chosen molecular representations and the way compound similarity is assessed.[2,3] The choice of molecular representations determines chemical reference spaces. For example, compound similarity relationships within a data set are expected to differ, dependent on whether the molecules are represented as different binary fingerprint vectors or arrays of numerical property descriptors. These different types of molecular descriptors yield distinct chemical reference spaces where given molecules might be more or less similar to each other. Hence, the topology of the corresponding activity landscapes is expected to change. Accordingly, different chemical space representations have been investigated for compound data sets and activity cliffs formed on the basis of different molecular representations have been compared,[10] giving rise to the notion of consensus activity cliffs, i.e., activity cliffs that are consistently observed when applying different molecular descriptors and chemical similarity methods.[10]

For the visualization of activity landscapes, 2D representations have thus far predominantly been used. Activity landscape representations originated with the introduction of structure−activity similarity (SAS) maps,[11] plots of structural similarity versus calculated activity similarity that delineate smooth landscape regions of high activity similarity and low structural similarity and rugged regions of high structural similarity and low activity similarity. In these plots, each data point represents a comparison of a pair of compounds in a data set. Prior to the introduction of SALI maps and NSGs, as discussed above, 2D similarity/potency correlation graphs were introduced[4] that are reminiscent of SAS maps but report 2D compound similarity relative to differences in potency and color-code compound pairs according to absolute potency values. These graphs were designed to compare 2D similarity and potency relationships of ligand sets, describe variable activity landscapes, and identify continuous and discontinuous SAR regions.[4] Another recent derivative of SAS maps are so-called multifusion similarity (MFS) maps[12] that utilize different compound 2D similarity measures and represent them following data fusion.

Although much information can be deduced from 2D representations of activity landscapes, 3D representations that are reminiscent of topographical maps are probably the most intuitive and elegant way of visualizing activity landscapes. Accordingly, this model has often been utilized to illustrate eminent features of activity landscapes, such as smooth regions and activity cliffs, and to rationalize conceptual relationships to continuous, discontinuous, and heterogeneous SARs.[1−3] However, although this idealized 3D landscape model has been widely discussed, actual 3D landscapes of compound data sets, i.e., "true" activity landscapes, have thus far not been described in detail.

Herein we present activity landscape representations of different types of compound sets that are calculated from potency data and pairwise compound distances in chemical space. A methodological framework is introduced for a consistent 3D approximation of activity landscapes of different compound sets. These representations are generated utilizing a conserved reference frame, which renders activity landscapes of different data sets directly comparable and makes it possible to study how different molecular representations might change the topology of landscapes. Visualization of 3D landscapes provides an intuitive access to prominent activity cliffs and the compounds that form them. In addition, activity landscapes of compound data sets having different characteristics according to SAR discontinuity score calculations can be compared.

## METHODOLOGY

**Activity Landscape Construction.** First we outline the approach to generate an activity landscape representation. For a given compound data set, 2D molecular graphs and potency measurements are required as basic input data. Figure 1a shows a schematic representation of a similarity/potency correlation graph as a prototypic 2D landscape visualization. For this landscape view, molecular representations are calculated from 2D graphs, and their similarity is calculated in a pairwise manner. Each data point represents a pairwise comparison yielding structural similarity and potency differences. In order to generate a 3D landscape representation with intuitive topological features, as schematically shown in Figure 1b, other types of calculations are required. For such a 3D representation, molecules must be projected into a 2D chemical reference space that is spanned by two molecular descriptors defining the *x*- and *y*-direction. These descriptors can be of a different type, for example, selected or combined contributions from molecular property descriptors or coordinates derived from molecular fingerprint similarity. A primary feature of 3D activity landscapes we need to capture are the activity cliffs that are formed by structurally similar molecules having dramatic potency differences. Figure 1c shows representative examples of compounds forming steep activity cliffs of large magnitude. Three-dimensional landscape design also starts with calculating molecular descriptors/representations. From a chosen molecular representation (herein different fingerprints are used), a coordinate-free chemical reference space is generated by calculation of pairwise compound distances (dissimilarities). The set of all pairwise distances defines this reference space. Then, multidimensional scaling[13] is used to project these molecules from the coordinate-free reference space onto an *x*/*y*-plane on the basis of their chemical dissimilarities. The *z*-axis reports the potency values of the molecules. In order to obtain a coherent potency surface that is required to obtain an interpretable landscape topology, we utilize a geostatistical technique termed Kriging[14] to interpolate between data points. The individual steps involved in 3D activity landscape generation are described in detail in the following sections.

**Compound Data Sets.** For our analysis, we assembled six classes of specific enzyme inhibitors with reported potency values from the MDDR.[15] As summarized in Table 1, these data sets include between 112 and 252 compounds. The compound sets were assembled to span different dissimilarity ranges, vary in their potency distributions and display different SAR characteristics (as further described below). In addition to these lead optimization sets, a high-throughput screening (HTS) hit set was taken from PubChem[16] that contained 2398 active compounds and had consistently lower potency ranges, hence resulting in a very low degree of SAR discontinuity (Table 1).

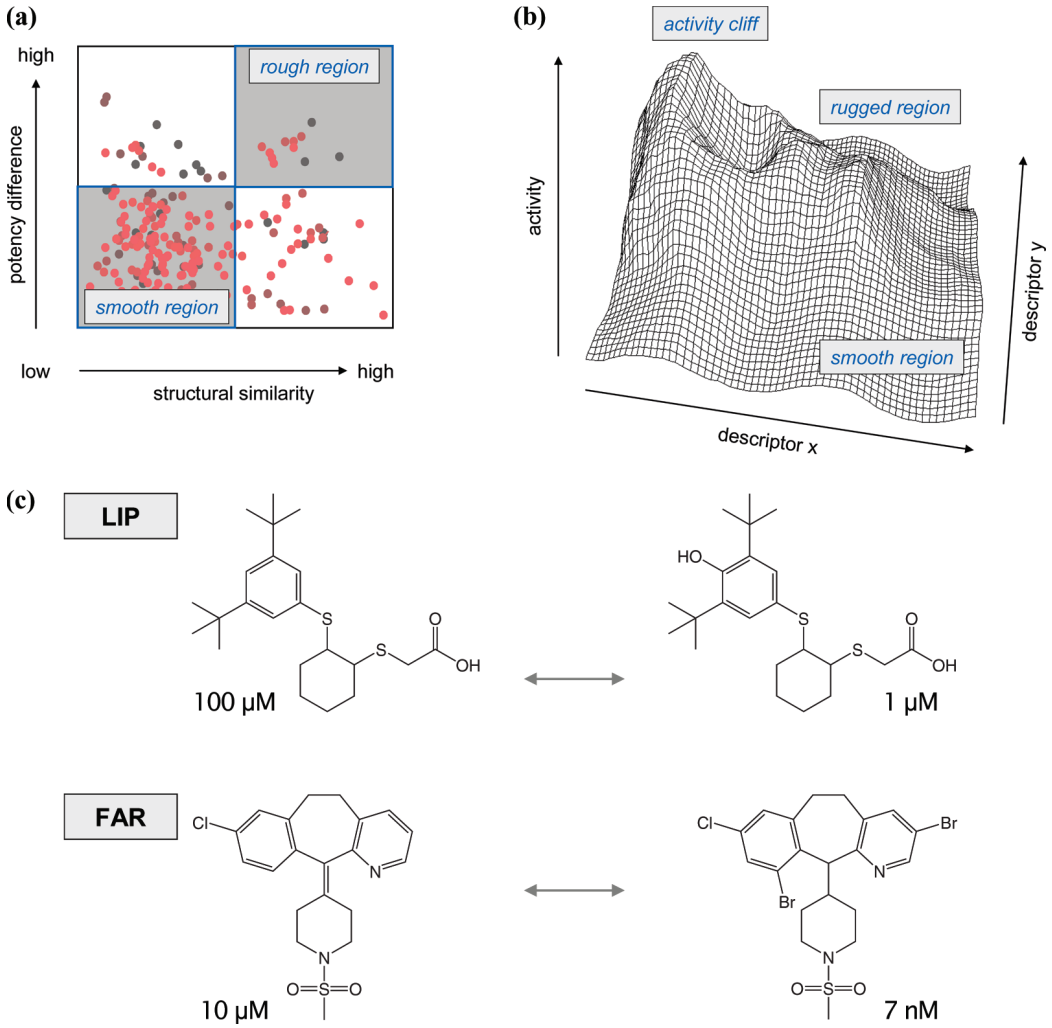



**Figure 1.** Schematic activity landscape representations and activity cliffs. (a) Similarity−potency plot. Pairwise structural similarity of active molecules is plotted against differences in logarithmic potency. Each data point represents a pairwise compound comparison and is colored according to the sum of the respective potency values using a continuous gradient from black for the lowest to red for the highest sum of potency values within a data set. Two characteristic regions are distinguished that contain pairs of molecules with low structural similarity and low potency difference, populating smooth regions of an activity landscape, or molecules with high structural similarity and large differences in potency, forming rough landscape regions. These regions contain activity cliffs. (b) Schematic 3D representation of an activity landscape. The *x*/*y*-plane represents a 2D projection of chemical space spanned by two descriptors that can be derived from different molecular representations, and the *z*-axis reports compound potency. The landscape contains idealized smooth and rugged (rough) regions and activity cliffs and hence corresponds to a heterogeneous SAR phenotype. (c) Examples of activity cliffs. Two exemplary compound pairs are shown from the LIP and FAR data sets, respectively, which have very similar structure but potency differences of several orders of magnitude and thus form activity cliffs of large magnitude.

**Table 1.** Summary of the Analyzed Enzyme Inhibitor Classes[a]

| | | | MACCS | | Molprint2D | | TGT | |
|---|---|---|---|---|---|---|---|---|
| activity class | no. of compounds | potency range | avg | max | avg | max | avg | /max |
| FAR | 146 | 3.52−10.54 | 6.33 | 9.22 | 7.01 | 8.83 | 14.05 | 23.39 |
| LIP | 252 | 4.00−9.00 | 6.56 | 9.11 | 6.03 | 8.25 | 12.28 | 19.80 |
| ACA | 195 | 3.92−9.59 | 6.16 | 8.83 | 6.17 | 8.94 | 12.02 | 20.86 |
| THR | 172 | 4.25−11.72 | 6.05 | 9.27 | 6.87 | 9.79 | 15.23 | 26.15 |
| ACH | 112 | 4.07−10.70 | 5.91 | 8.72 | 6.06 | 8.00 | 11.30 | 18.57 |
| 5HT | 129 | 5.57−11.00 | 5.68 | 8.54 | 6.06 | 7.94 | 11.36 | 20.03 |
| HADH2 | 2398 | 4.40−7.60 | 6.53 | 9.49 | 6.00 | 8.60 | 12.04 | 23.17 |

[a] For the seven compound activity classes discussed in the text, the number of compounds, potency range, and average (avg) and maximum (max) Euclidean fingerprint distances are reported. The minimum distance was 0 for all classes and fingerprint representations. Activity classes are abbreviated as follows: protein farnesyltransferase inhibitors (FAR), lipoxygenase inhibitors (LIP), acyl-CoA:cholesterol acyltransferase inhibitors (ACA), thrombin inhibitors (THR), acetylcholinesterase inhibitors (ACH), 5HT reuptake inhibitors (5HT), and human hydroxyacyl−CoA dehydrogenase II (PubChem BioAssay ID 886).

**Molecular Representation.** Test compounds are initially projected into a low-dimensional chemical reference space. For this purpose, we define a coordinate-free reference space based on Euclidean distances between molecular fingerprint representations. Three conceptually different fingerprint designs are applied: MACCS,[17] TGT,[18] and Molprint2D.[19]

MACCS is a widely used structural key-type fingerprint that monitors the presence or absence of predefined structural features in a molecule. With 166 bit positions corresponding to 166 distinct structural features, its structural "resolution" is relatively low. By contrast, TGT represents a topological three-point pharmacophore fingerprint that monitors all triplets of predefined pharmacophore features with a given bond distance in a molecule and consists of 1704 bits. Molprint2D captures layered atom environments as a measure of the global topology of a molecule. Because it does not rely on a catalogue of predefined substructures, its format is flexible, and Molprint2D can generate a theoretically unlimited number of features for a molecule. Thus, this fingerprint representation is of high structural resolution.

**Chemical Dissimilarity Assessment.** A variety of similarity or distance measures are available for the comparison of molecular fingerprints.[20] In this study, the dissimilarity of two molecules is calculated as the Euclidean distance between their fingerprint representations. For binary fingerprints, the Euclidean distance is defined as follows:

$$\delta_{ij} = \sqrt{N_i + N_j - 2N_{ij}}$$

where $N_i$ and $N_j$ denote the number of fingerprint features present in molecules $i$ and $j$, respectively, and $N_{ij}$ denotes the number of features shared by both molecules. The Euclidean distance is chosen here instead of the widely applied Tanimoto similarity coefficient[20] for two reasons. First, the Tanimoto coefficient is calculated only on the basis of features that are present in two molecules and does not account for features that are absent. By contrast, the Euclidean distance calculates molecular dissimilarity on the basis of features that differ between two molecules. For the purpose of landscape visualization, we found that simple Euclidean distance calculations often better differentiated between similar molecules than those of Tanimoto similarity calculations, which is relevant with respect to data spread and surface coverage. However, landscapes produced on the basis of Tanimoto similarity and Euclidean distances were often rather similar, suggesting that Tanimoto similarity could also be utilized. Nevertheless, for our purposes, Euclidian distance has a second principal advantage because it provides a standard framework for the comparison of numerical molecular descriptors, which might also be used for landscape generation, as an alternative to fingerprints.

**Reference Space Construction.** For computational analysis, molecules are generally projected into a chemical reference space that is defined by a set of molecular descriptors or fingerprint vectors. Reference spaces are typically high-dimensional and hence difficult to represent in an intuitive and readily interpretable manner. To enable the visualization of chemical space distributions of large molecular data sets, various dimensionality reduction techniques have been introduced that aim at mapping multidimensional data into 2D or 3D reference spaces.[21] These reference spaces can either be coordinate-based or coordinate-free, depending on the dimension reduction method that is used. One of the most common techniques is principal component analysis (PCA) that generates a low-dimensional coordinate-based space from linear combinations of original descriptors with minimal loss of data variance.[22] An advantage of this method is that novel molecules can easily be mapped into principal components space. This provides the basis for the ChemGPS method[23] that utilizes principal components precalculated on a set of active compounds to generate coordinates of novel input molecules. By contrast, methods like nonlinear mapping (NLM)[24] or multidimensional scaling (MDS)[13] aim at preserving relative similarity relationships between input data points by minimizing a stress function (see below) and thus produce coordinate-free low-dimensional reference spaces. These methods often reflect close similarity relationships better than coordinate-dependent approaches. However, they are computationally demanding and not easily applicable to large data sets. This problem can be overcome, for example, by combining MDS with artificial neural networks.[25] Another alternative is presented by Kohonen networks that project data onto a 2D map using a self-organizing learning algorithm.[26]

Here we apply a nonmetric multidimensional scaling algorithm to visualize molecular dissimilarity relationships. For a set of $n$ molecules, the algorithm takes as input an $n$x$n$ matrix of pairwise Euclidean distances $\delta_{ij}$ of molecular fingerprints, as defined above, and calculates $n$ points with 2D coordinates $(x_i,y_i)$, whose pairwise Euclidean distances $d_{ij}$ best approximate the input dissimilarities $\delta_{ij}$. Specifically, we aim to find $n$ 2D vectors $p_i = (x_i,y_i)$ such that Kruskal's stress function[27] is minimal:

$$\text{stress} = \sqrt{\frac{\sum_{i<j} (d_{ij} - \hat{\delta}_{ij})^2}{\sum_{i<j} d_{ij}^2}}$$

where $d_{ij}$ denotes the Euclidean distance between points $p_i$ and $p_j$:

$$d_{ij} = d(p_i, p_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

and $\hat{\delta}_{ij}$ denotes an optimal monotonic transformation of the input dissimilarities $\delta_{ij}$ that is determined by the optimization algorithm.[28] The optimization problem is solved by means of an iterative steepest-descent algorithm implemented in the "MASS" package[29] of R.[30] The resulting coordinates assigned to each molecule are then scaled to the range [0,1] by subtracting the minimum and dividing by the range of the $x$- and $y$-values. Subsequently, the scaled coordinates are multiplied by the maximal chemical dissimilarity between two molecules in the current data set. Thus, the range of the planar coordinates (and hence the size of the landscape plots) reflects the overall chemical dissimilarity within a data set.

**Surface Interpolation.** Multidimensional scaling generates an embedding of active molecules in a 2D plane. Potency values are then added as the third dimension for the activity landscape model. In general, however, the data points are sparse and unevenly distributed and must be interpolated to obtain a coherent surface. For this purpose, a geostatistical technique termed Kriging[14] is applied to fit a coherent surface to the data points. This method aims at estimating the value of a random field, in our case the surface elevation, at unobserved locations from observations at $n$ data points, i.e., the $n$ given molecules with their position on the $x$/$y$-plane and their potency value on the $z$-axis. Based on the expected value and a covariance function that describes the spatial dependence of the given data points, the Kriging method

**Table 2.** Evaluation of the Interpolated Activity Landscapes[a]

| | correlation between chemical and geometric distances | | | correlation between interpolated and original potency values | | | percentage of interpolated surface area | | |
|---|---|---|---|---|---|---|---|---|---|
| activity class | MACCS | M2D | TGT | MACCS | M2D | TGT | MACCS | M2D | TGT |
| FAR | 0.73 | 0.51 | 0.81 | 0.98 | 0.96 | 0.85 | 23.2 | 28.7 | 27.7 |
| LIP | 0.75 | 0.71 | 0.68 | 0.97 | 0.92 | 0.88 | 6.0 | 10.4 | 20.4 |
| ACA | 0.78 | 0.80 | 0.79 | 0.96 | 0.92 | 0.94 | 12.3 | 7.7 | 15.7 |
| THR | 0.69 | 0.50 | 0.76 | 0.93 | 0.93 | 0.92 | 20.5 | 17.3 | 9.9 |
| ACH | 0.81 | 0.60 | 0.74 | 0.98 | 0.97 | 0.96 | 14.8 | 18.1 | 19.2 |
| 5HT | 0.81 | 0.73 | 0.81 | 0.96 | 0.97 | 0.94 | 17.1 | 15.1 | 25.7 |
| HADH2 | 0.55 | 0.27 | 0.69 | 0.77 | 0.66 | 0.61 | 6.8 | 9.1 | 13.7 |

[a] For the three fingerprint representations, MACCS, Molprint2D (M2D), and TGT, correlations between calculated Euclidean fingerprint distances (chemical distances), and geometric distances between 2D molecular coordinates obtained by multidimensional scaling are reported. Furthermore, correlations between the interpolated surface values and the original potency values are provided. In addition, the percentage of grid points that are displayed fully transparent (white) and represent purely interpolated surface area is given (see text for details).

calculates the best linear unbiased estimator for the surface elevation by minimizing the variance of the prediction error. The surface is calculated on a regular grid consisting of 80 × 80 grid points. Because the molecules are in most cases not evenly distributed on this grid, border regions occur where no data points are present to support the interpolation. These regions are omitted in the landscape plots, which can sometimes result in irregularly shaped borders of the images. We utilize the Kriging function as implemented in the "fields" package of R.[31]

**Graphical Display.** The resulting activity landscapes are displayed as perspective plots generated with R. To enable the comparison of landscapes across different activity classes and fingerprint representations, all landscape representations have been generated from the same viewpoint (i.e., with an azimuth of 45° and a colatitude of 25°). Moreover, a common scale for the *z*-axis is applied for all data sets, ranging from the lowest (3.72) to the highest (11.55) interpolated *z*-values observed for all six MDDR activity classes. In addition, for each fingerprint representation, a common scale is utilized on the *x*- and *y*-axes to make the landscapes for a given fingerprint comparable to each other. This scale ranges from the lowest (0.00) to the highest values of chemical distances for the respective fingerprints over all six MDDR classes (MACCS − 9.27, Molprint2D − 9.79, and TGT − 26.15). The surface facets are colored according to *z*-values. Areas with a *z*-value below a lower threshold of 5.78 are colored in green, and areas with a *z*-value above an upper threshold of 8.75 are colored in red. These threshold values are determined as the highest minimal and the lowest maximal *z*-values of the six MDDR activity classes, respectively, and make it possible to directly identify regions in a landscape where interpolated potency values are above or below a given value, which might be difficult to recognize on the basis of surface elevation alone. Intermediate values are colored using a continuous gradient from green via yellow to red. For the HTS data, we set the thresholds for green and red coloring to 4 and 7, respectively, in order to account for the narrow potency range and the presence of large numbers of only very weakly active molecules in this compound set. In addition, coloring is designed to convey information about the data sampling of the surface: colors fade with increasing distance of a surface facet to a data point; hence, white areas denote regions that are not populated by data points and represent interpolated surface areas. The transparency (α) value of each grid point *p* is determined from the Euclidean distance $d(p,(x_i, y_i))$ of *p* to the closest data point $(x_i, y_i)$, representing the coordinates of a molecule *i* calculated by multidimensional scaling:

$$\alpha(p) = 255 - \min_i\{d(p,(x_i,y_i))\} \cdot \frac{k}{x\text{max} - x\text{min}}$$

Here, *x*max and *x*min denote the largest and smallest *x*-coordinates of the landscape area, and *k* is a scaling factor that determines the slope of the transparency gradient. In our calculations, *k* was empirically set to 1800. With this formulation, grid points that map close to a data point obtain α values near 255, which corresponds to an opaque coloring, whereas grid points whose distance to the closest data point is large obtain low α values near 0, which results in fully transparent (or white) representation. Negative α values are set to 0. It follows from the equation that grid points whose distance to the nearest data point is $(255)/(k)(x\text{max} - x\text{min})$ or larger will obtain a minimal transparency value of 0 and are displayed in white; these grid points form purely interpolated surface areas. The percentage of these grid points is reported in Table 2 for each activity class and for all three fingerprint representations, which provides a quantitative comparison of the landscape representations.

**SAR Discontinuity Scores.** To quantify the presence of activity cliffs in a compound data set, we calculate the SARI discontinuity score.[4,7] This score has been introduced to estimate the global SAR character of an activity class A and computes the average potency difference between pairs of similar compounds, scaled by pairwise similarity:

$$\text{disc}_{\text{raw}}(\text{A}) = \underset{\left\{\begin{matrix}(i,j)\in \text{A}|\delta_{ij}<t,\\ |P_i-P_j|>1\end{matrix}\right\}}{\text{mean}} \left(|P_i-P_j| / 1+\delta_{ij}\right)$$

Here, $P_i$ denotes the negative decadic logarithm of the potency value of compound *i*, and $\delta_{ij}$ is the Euclidean fingerprint distance of compounds *i* and *j*; *t* denotes a fingerprint distance threshold that was set to 4.90 for MACCS, 8.31 for TGT, and 5.29 for Molprint2D. These values were chosen to eliminate the same percentage (9.24%) of pairwise compound distances from a set of 13 reference classes originally used for MACCS Tc calculations.[7] The global discontinuity scores for each activity class and fingerprint combination are given in Table 3. In addition, Table 3 also reports the number of activity cliff markers in landscapes that correspond to individual compounds partici-

**Table 3.** Discontinuity Scores and Activity Cliffs[a]

| | discontinuity score | | | no. of activity cliff markers | | |
|---|---|---|---|---|---|---|
| activity class | MACCS | M2D | TGT | MACCS | M2D | TGT |
| FAR | 0.79 | 0.64 | 0.77 | 39 (26.7%) | 13 (8.9%) | 30 (20.5%) |
| LIP | 0.09 | 0.04 | 0.14 | 8 (3.2%) | 11 (4.4%) | 12 (4.8%) |
| ACA | 0.23 | 0.34 | 0.18 | 24 (12.3%) | 45 (23.1%) | 20 (10.3%) |
| THR | 0.59 | 0.69 | 0.56 | 71 (41.3%) | 25 (14.5%) | 7 (4.1%) |
| ACH | 0.75 | 0.83 | 0.64 | 48 (42.9%) | 41 (36.6%) | 30 (26.8%) |
| 5HT | 0.24 | 0.33 | 0.27 | 24 (18.6%) | 21 (16.3%) | 18 (13.9%) |
| HADH2 | 0.05 | 0.06 | 0.07 | 48 (2.0%) | 452 (18.8%) | 37 (1.5%) |

[a] SARI discontinuity scores calculated on the basis of Euclidean distance between MACCS, Molprint2D (M2D), and TGT fingerprints are reported for the seven compound activity classes. In addition, we report the number and percentage (in parentheses) of "activity cliff markers", i.e., molecules that participate in at least one compound pair with fingerprint distance that is lower than the distance threshold applied for discontinuity score calculations and potency differences of more than three orders of magnitude.

pating in at least one compound pair with fingerprint distance less than the threshold specified above and the potency differences of at least 3 orders of magnitude. If such compound pairs are proximal on an activity landscape, then they participate in the formation of an activity cliff region consisting of multiple and in part overlapping cliffs.

**Compound Clustering.** In order to enable a detailed analysis of compound classes forming different parts of activity landscapes, in particular, activity cliffs, we also clustered the molecules in a data set on the basis of pairwise Euclidean fingerprint distances. For this purpose, the hierarchical clustering scheme of Ward's minimum-variance linkage method was applied.[32] The resulting dendrograms were pruned at various heights to obtain a reasonable number of clusters with balanced cluster composition. We also calculated the discontinuity score for each resulting cluster to evaluate local SAR features that might coexist within a given data set. Cluster results for all seven activity classes are provided in the Supporting Information.

The landscape display and analysis tools introduced herein enable rotatable landscape views, molecule selection, and interactive structure display. Upon publication, these tools are made freely available via the following: http://www.lifescienceinformatics.uni-bonn.de.

## RESULTS AND DISCUSSION

**Landscape Generation and Interpretation.** We have generated both 2D and 3D activity landscape models for seven enzyme inhibitor sets, including six compound optimization sets and one screening set, using three different molecular fingerprint representations. Figure 2a shows a representative example for the ACH data set and MACCS fingerprints that is utilized to rationalize key features of landscapes revealed by our analysis and to illustrate how 3D landscape representations should be interpreted in order to identify key compounds. In the 2D representation of the ACH landscape, molecules are represented by data points whose coordinates were obtained by multidimensional scaling, as used for the generation of the 3D landscape representation. The interpolated surface elevation is represented by shading, using the same color code as in the 3D landscape. Corresponding exemplary data points in the 2D and 3D representations are connected by dashed lines. The 2D landscape representation is intuitive and mirrors the data distribution, but the 3D landscape further emphasizes the formation of activity cliffs and their spatial arrangement. Only three major analysis criteria must be applied, as indicated on the left in Figure 2a, to interpret activity landscapes in a step-by-step manner, to evaluate characteristic landscape features, and to focus on key compounds:

(i) Regions of interpolated surface area (white) are identified that are particularly "smooth" but lack compound data. These regions contribute to landscape topology but lack interpretable local SAR information. Hence, this information can be utilized to assess the sampling of a compound data set and to identify chemical space regions that have not been thoroughly explored.

(ii) Regions with green to yellow peaks of limited magnitude are then identified that result from dense data sampling but do not correspond to local regions of significant SAR discontinuity, as we discuss in more detail below. Therefore, these moderate surface elevations are termed "data peaks". This is an important point to be made because not every peak on a 3D landscape represents an activity cliff.

(iii) True activity cliffs become immediately apparent on a 3D landscape in regions of large-magnitude peaks that are characterized by a red−yellow−green color spectrum. These peaks are formed by groups of similar molecules that map close to each other in the reference space but have distinct potency levels. Hence, to identify prominent activity cliffs, color-code information, indicating absolute potency differences among similar molecules, must be taken into account, as is also further discussed below.

In Figure 2b, the results of compound clustering and landscape mapping are shown, revealing that different chemotypes form spatially separated activity cliffs in the ACH data set, as one would expect. The individual clusters obtain discontinuity scores that span the entire range from 0 to 1, which indicates the coexistence of different local SAR features within the compound set. Molecules belonging to two clusters characterized by a notable degree of SAR discontinuity are mapped on the 3D landscape view in Figure 2b, and the structures of two compound pairs forming prominent activity cliffs are shown. Furthermore, representative data points that correspond to the most active compounds in each cluster are displayed on the 3D surface in Figure 2b, and their structures are shown in Figure 2c. These molecules represent different chemotypes and produce distinct peaks in the activity landscape that are scattered around the surface area. Similar observations were made for all seven compound data sets, as shown in Supporting Information, Figure S1.

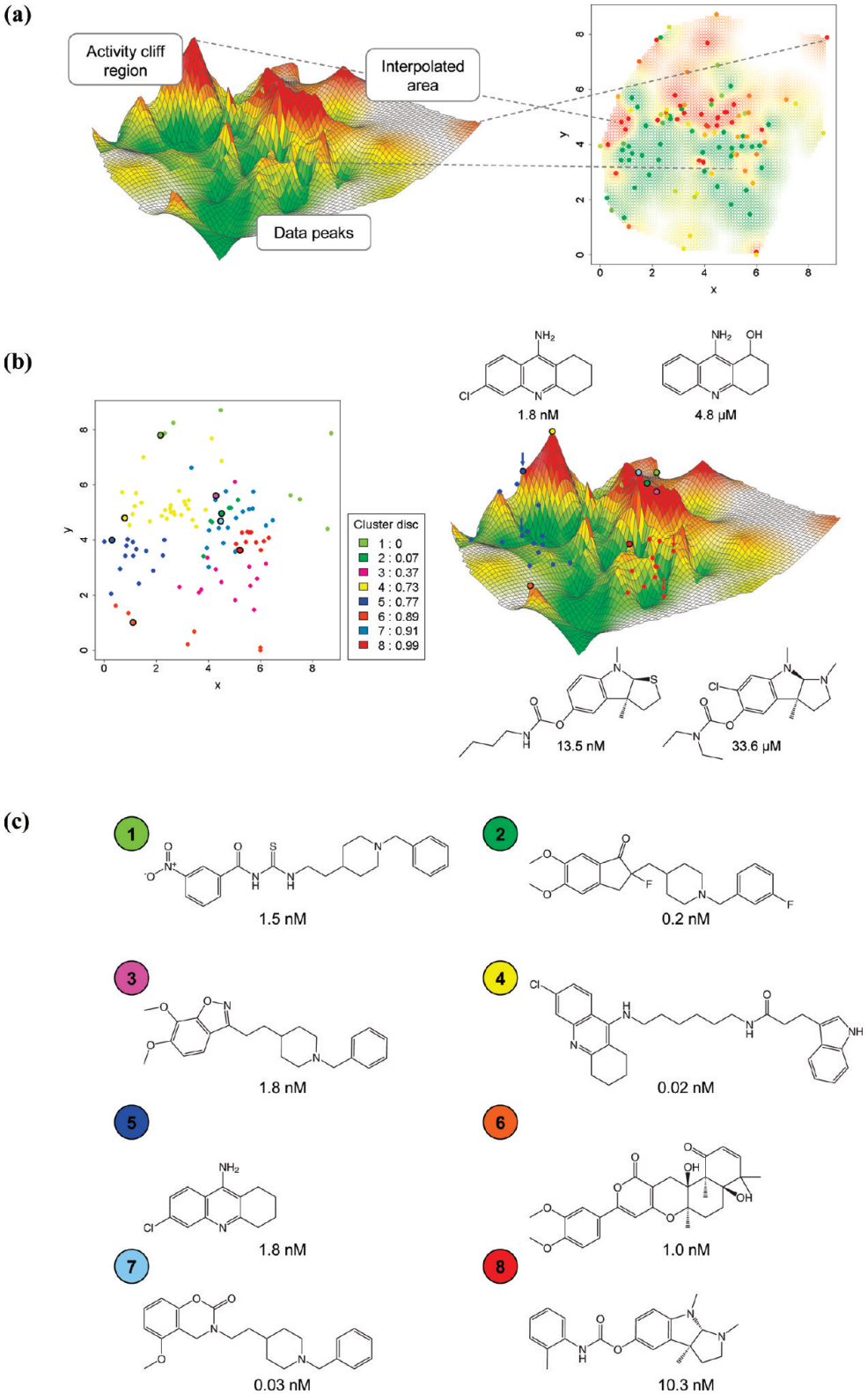



**Figure 2.** Interpretation of activity landscape representations. For the ACH data set and MACCS fingerprints, 2D and 3D activity landscape representations are shown. (a) Comparison of 2D and 3D landscape. The 3D landscape (left) contains distinct regions that are discussed in the text. These regions can be mapped onto a 2D representation of the same landscape (right) obtained by multidimensional scaling. In the 2D plot, the interpolated surface elevation is represented by shading, using the same color scheme as in the 3D landscape. Data points representing molecules are also shown and colored according to their potency values, with green indicating potency values of 5.78 and below and red indicating potency values of 8.75 and above. (b) Cluster analysis. The compounds in the data set were clustered using Ward's hierarchical clustering based on Euclidean fingerprint distances. In the 2D plot (left), data points representing molecules are colored according to their cluster membership. SARI discontinuity scores calculated for each cluster are in the box ("Cluster disc"). The most active compound in each cluster is encircled and also shown on the 3D landscape (right). In addition, two clusters are mapped onto the 3D landscape. (c) Cluster representatives. Shown are the structures of the most potent compounds in each cluster marked in (b).

**Landscape Quality Assessment.** The six lead optimization sets produced characteristic 3D landscape topologies that differed in part substantially depending on the choice of the molecular representation. These differences are discussed

below in detail. In order to evaluate the overall quality of the models, we compared the modeled parameters for molecular distance and surface elevation to the chemical descriptor distance and measured potency data, respectively. The correlation values are reported in Table 2. For distance comparison, we calculated the pairwise Euclidean distances between molecule coordinates obtained through multidimensional scaling and correlated these geometric distances to the Euclidean fingerprint distances. On average, geometric and fingerprint distances correlated well (0.72) and exceeded a correlation of 0.6, with the exception of only 2 of 18 compound class/fingerprint combinations (Molprint2D for classes FAR and THR). However, geometric distances calculated with a conventional multidimensional scaling algorithm[33] displayed consistently lower correlation with fingerprint distances, which supported our choice of a nonmetric approach to multidimensional scaling.

Comparison of interpolated surface elevation with measured potency values yielded correlations that were greater than 0.85 for all activity class/fingerprint combinations (and exceeded 0.9, except for FAR and LIP with TGT fingerprints). Hence, according to parameter correlation analysis, the 3D activity landscape models were generally of good quality. Importantly, all activity landscapes studied here were generated using a consistent data reference frame that made it possible to compare landscapes across different activity classes.

**Global SAR Features of Lead Optimization Sets.** The SARI discontinuity scores reported in Table 3 are a global measure of SAR characteristics. Discontinuity scores range from 0 to 1. The higher the discontinuity score is the more structurally similar compounds with significant potency differences are contained in a data set (and the more activity cliffs are formed). By contrast, low discontinuity scores indicate the presence of only small potency differences among structurally dissimilar compounds and the absence of activity cliffs of large magnitude. Hence, these global discontinuity scores should correlate with notable differences in landscape topology. The scores were calculated with three different fingerprints. As can be seen in Table 3, the values differ in each case but are comparable in magnitude for each class, indicating the presence of high SAR discontinuity for the activity classes farnesyltransferase (FAR) and acetylcholinesterase (ACH) inhibitors, intermediate discontinuity for thrombin (THR) inhibitors, and low discontinuity for inhibitors of lipoxygenase (LIP), acyl-CoA:cholesterol acyltransferase (ACA), and 5HT reuptake (5HT). Thus, these activity classes cover a wide range of SAR discontinuity. Table 3 also lists the number of prominent activity cliffs contained in each compound set.

**Landscape Topology and Molecular Representations.** The calculated FAR activity landscapes in Figure 3a clearly reflect the high degree of SAR discontinuity contained in this data set, which is particularly well illustrated by the landscape calculated with Molprint2D. Here, compounds are distributed over the entire landscape, resulting in the presence of only small interpolated (white) surface regions. The landscape is rugged and characterized by multiple cliffs, some of which are not separated and form a plateau of highly potent compounds (coherent red region). The MACCS- and TGT-based landscapes also display a rugged topology. Different from the landscape calculated with Molprint2D, the MACCS-based landscape is characterized by a large interpolated surface area, which is a consequence of clear separation of highly (red areas) and weakly potent (green) compounds. Similarly, the TGT-based landscape also contains a large interpolated surface area, but the topology of this landscape differs substantially from the others. This is the case because the calculation of TGT pharmacophore feature fingerprints results in clustering of different compound subsets, rather than a separation of molecules according to potency. Thus, the comparison of the three FAR landscapes illustrates a strong influence of the chosen molecular representation on landscape topology, although all three landscapes capture the high degree of SAR discontinuity within the FAR data set well. Similar observations can be made for all activity landscapes studied here, as discussed in the following.

**SAR Discontinuity versus Continuity.** Comparison of activity landscapes for the different compound sets shows that they all include a number of peaks and rugged regions, despite differences in global SAR character. For example, the LIP data set is characterized by a very low discontinuity score for all three fingerprints. Inspecting its activity landscapes, shown in Figure 3b, reveals that this large data set evenly populates the landscapes, except for the TGT representation where clustering effects also occur in this case. The MACCS- and Molprint2D-based landscapes are rather similar, despite minor differences in topology. In these landscapes that are dominated by moderately potent molecules (green and yellow areas) prominent cliffs are absent; however, many small peaks are scattered over the surface. It should be noted, however, that these peaks primarily result from the underlying data point distribution and are in this case not indicative of SAR discontinuity. This is the case because their height is rather limited and they are mostly colored in similar green and yellow shades, which indicates that the corresponding molecules have similarly weak potency values and do not form activity cliffs. As illustrated in the bottom part of Figure 3b, removing the 30 and 100 most active molecules from the LIP data set makes these landscapes smoother. However, even after removal of 100 molecules (which limits logarithmic potency to the range between 6.9 and 9), the landscape still contains a number of small peaks. Hence, these peaks represent molecules whose potency is only slightly higher than that of its neighbors. By contrast, the classes FAR or ACH (see below) are characterized by a high discontinuity score, and accordingly, their landscapes contain rugged regions where peaks colored in red that are formed by highly potent molecules are in close proximity to valleys or canyons where weakly active molecules are located. Thus, in order to detect SAR discontinuity and activity cliffs in a 3D activity landscape, the height and color of neighboring peaks and valleys must be taken into account.

Similar to LIP, the ACA data set also contains many weakly to moderately potent compounds but is characterized by a higher degree of discontinuity, which becomes apparent in its activity landscapes shown in Figure 3c. Here the compounds are also well distributed over most of the surface areas, but the landscapes consist of different regions that are predominantly populated by either weakly or moderately to highly potent compounds. In the latter regions, small- to moderate-sized activity cliffs are formed.

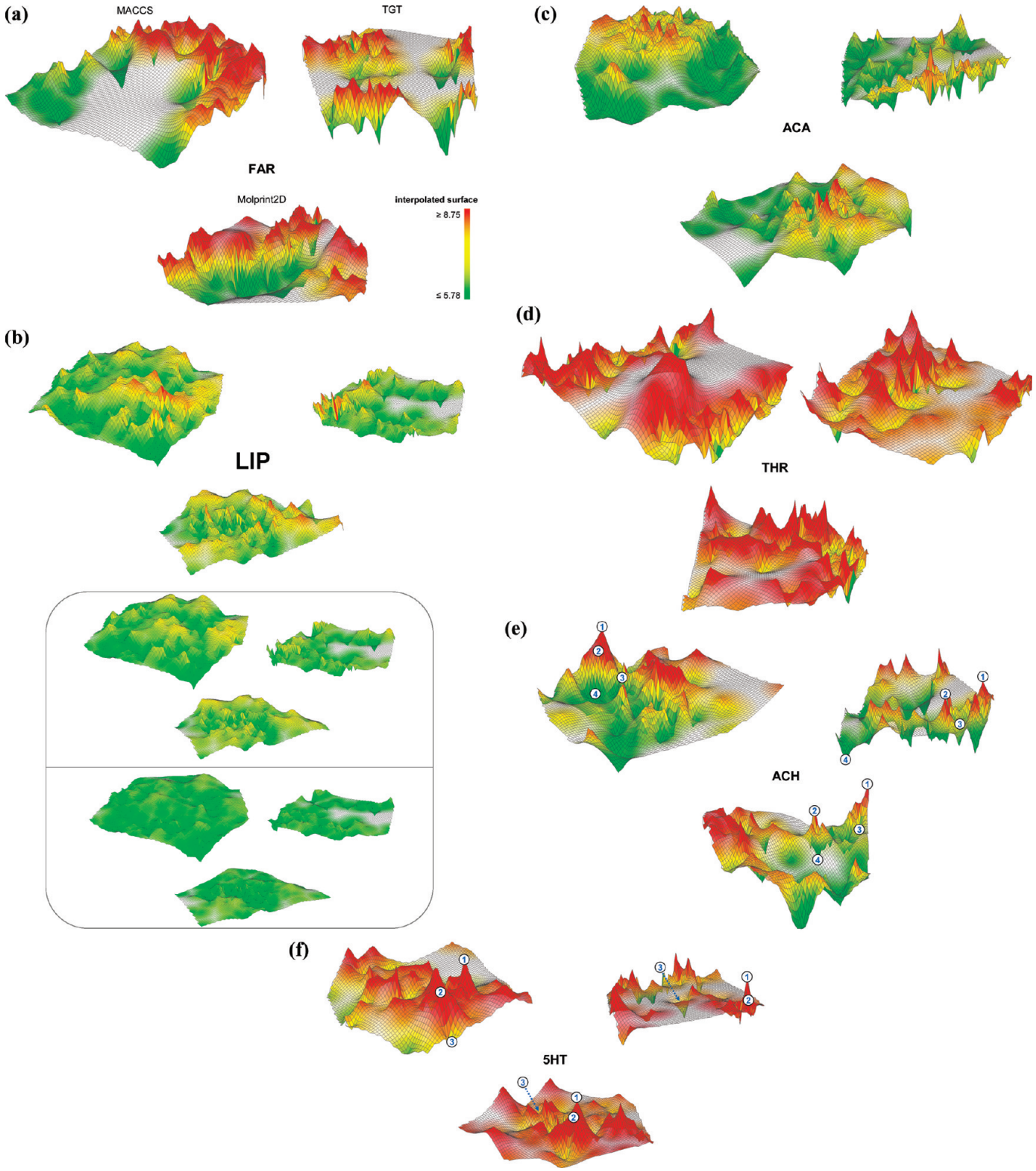



**Figure 3.** Activity landscapes. For the six compound data sets according to Table 1, activity landscapes were calculated on the basis of Euclidean fingerprint distances for three fingerprint representations, MACCS (top left), TGT (top right), and Molprint2D (bottom). The surface is colored according to interpolated surface elevation, using a continuous spectrum from green for values smaller than or equal to 5.78 to red for values equal to or greater than 8.75. For all combinations of the six activity classes and three fingerprints, the same color spectrum and a common coordinate reference frame are applied. Interpolated surface area not populated with molecules is colored white. Activity landscape representations are shown for inhibitors of: (a) protein farnesyltransferase (FAR), (b) lipoxygenase (LIP), (c) acyl-CoA: cholesterol acyltransferase (ACA), (d) thrombin (THR), (e) acetylcholinesterase (ACH), and (f) 5HT reuptake (5HT). The box in the lower part of Figure 3b shows activity landscape representations for class LIP that were calculated after removal of the 30 (top) and 100 (bottom) most active compounds from the data set. Relatively high peaks are smoothed out in the resulting landscapes, but small peaks are retained. The comparison of these landscapes illustrates the effect of data sampling and the difference between peaks produced by dense data points and actual activity cliffs (see text for details).

Different from LIP and ACA, the THR inhibitor set is dominated by highly potent compounds. It yields intermediate discontinuity scores that indicate SAR heterogeneity, which usually results from the presence of subsets of

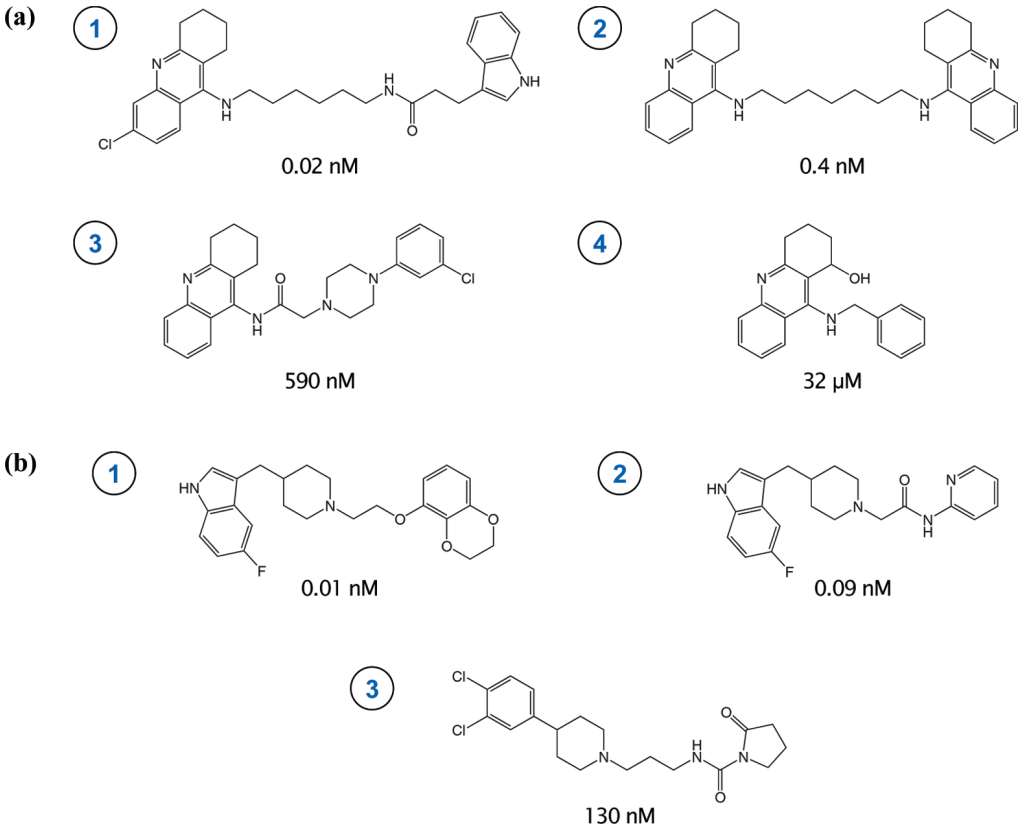



**Figure 4.** Exemplary compounds. For (a) ACH and (b) 5HT molecules are shown that are labeled in the activity landscapes in Figure 2e and Figure 2f, respectively. Depending on the chosen fingerprint representation, these molecules map to different regions of the landscapes and form, or do not form, activity cliffs.

compounds displaying different SAR characteristics. Given the potency distribution within this compound set, its activity landscapes, shown in Figure 3d, predominantly consist of red and yellow regions. Here differences in landscape topologies produced by different fingerprints are again rather obvious, and depending on the fingerprint, different clustering patterns are observed. Although the MACCS- and TGT-based landscapes contain extended regions of interpolated surface, all three landscapes are characterized, despite topology differences, by smooth and relatively flat regions and also by regions that are enriched with cliffs of varying magnitude. The Molprint2D-based landscape has compounds distributed over most of its surface, and best reflects these features that are consistent with SAR heterogeneity. Taken together, these findings illustrate that the topological details of the individual activity landscapes of the four compound data sets discussed so far are much influenced by the different molecular representations. However, the results also show that compound set characteristic features common to these four activity landscapes are consistent with global SAR phenotypes assigned on the basis of discontinuity scoring.

**Variable Activity Cliffs.** Activity cliffs represent the most informative and characteristic features of activity landscapes. Consistent with the previously observed predominance of SAR heterogeneity in many compound data sets,[4,7] we find that essentially all activity landscapes, except those representing the most continuous SARs, contain activity cliffs of varying magnitude.

Consistent with high discontinuity scores for all three fingerprint representations, the landscapes for the ACH data set, shown in Figure 3e, are dominated by pronounced activity cliffs that are formed by compounds covering a large potency range from subnanomolar to micromolar potencies. However, the distribution of these cliff marker compounds differs substantially in the three landscapes, depending on the chosen fingerprint representation. Figure 4a shows four exemplary molecules representing different potency levels, whose positions on the landscapes in Figure 3e are indicated. These molecules share a common tricyclic substructure and mark activity cliffs. In the MACCS-based landscape, they map to the same surface area that contains a prominent activity cliff. The two highly potent molecules 1 and 2 contribute to a peak that is produced by a number of similarly potent molecules that map to this surface region. By contrast, the other two fingerprint representations clearly separate these compounds. In the Molprint2D-based landscape, the molecule pairs 1−3 and 2−4 form two separate activity cliffs of similar magnitude. By contrast, in the TGT-based landscape, the least potent (and smallest) molecule 4 maps to a different area distant from the location of the other three selected molecules. Hence, the formation of activity cliffs also varies with chosen molecular representations, more so than overall landscape topology.

The 5HT data set is characterized by a lower discontinuity score than ACH, which is due to the prevalence of highly potent compounds in the 5HT set. The 5HT activity landscapes in Figure 3f also include moderately sized activity cliffs that are formed by neighboring molecules with high and moderate (and, in a few cases, low) potency levels. Three exemplary molecules are labeled in Figure 3f and shown in Figure 4b. Molecules 1 and 2 are structurally very similar and located close to each other in all three activity landscapes, producing the highest peaks. Compound 3 is four to five orders of magnitude less potent than these two compounds

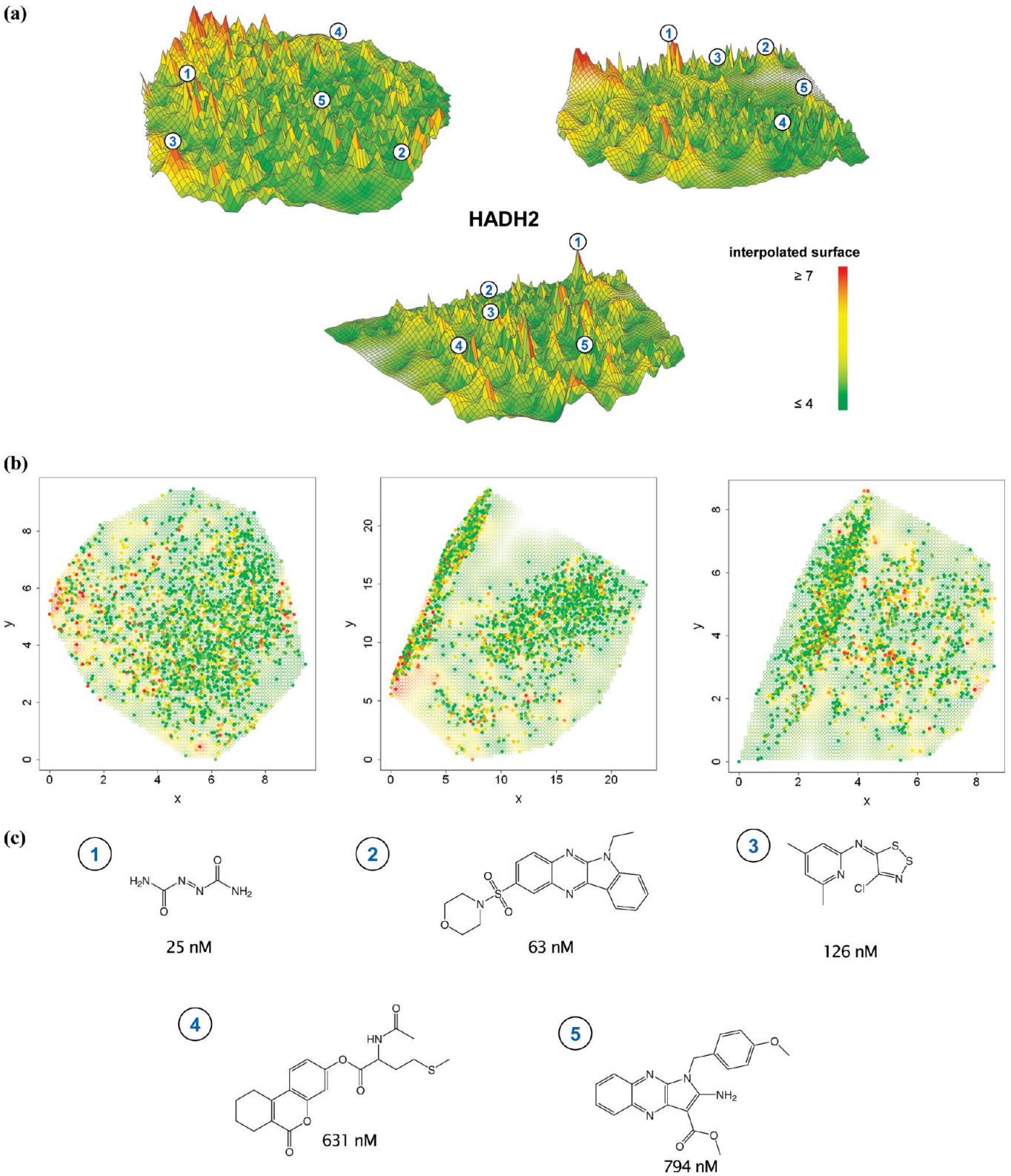



**Figure 5.** Activity landscape for HTS data. Activity landscape representations for a set of 2398 inhibitors of hydroxyacyl−CoA dehydrogenase II taken from a screening set are shown for three different fingerprint representations. (a) 3D landscape representations for MACCS (left), TGT (right), and Molprint2D (bottom) fingerprints. Representative molecules belonging to different clusters are indicated on the surface and are colored according to cluster membership. (b) 2D representations of the same activity landscapes, arranged according to (a). (c) Representative molecules belonging to different clusters marked in (a) are shown together with their potency values.

and structurally distinct from them. However, due to the presence of a common substructure, all three molecules map proximal to each other in a contiguous region in the MACCS-based landscape. By contrast, the other two higher-resolution fingerprint representations clearly separate compound 3 from the two highly potent molecules and place it into a more distant region in the corresponding activity landscapes. In this case, the higher-resolution fingerprints further emphasize activity cliffs and separate them on their activity landscapes.

**Activity Landscape Analysis of Screening Data.** In addition to compound optimization sets, we have also analyzed HTS data, given their relevance for initial SAR

exploration and hit selection. Screening data sets generally present challenging cases for systematic SAR analysis because their potency and similarity distributions differ substantially from compound optimization sets. To account for the narrow potency range, the color code applied for the 3D landscape representations has been modified: green coloring now corresponds to an interpolated surface elevation of 4 and lower, whereas red indicates a surface elevation of 7 and higher. This modification makes it possible to evaluate small potency differences in the data set (but the landscape coloring cannot be directly compared to the six MDDR data sets). The hydroxyacyl−CoA dehydrogenase II (HADH2) data set is characterized by the presence of many weakly or borderline active molecules that dominate its SAR character and lead to a very low degree of SAR discontinuity. Its activity landscape representations, shown in Figure 5a, clearly reflect this SAR phenotype. Many small green data peaks are seen that arise from dense data sampling. As a consequence of data density, purely interpolated surface area (represented as white regions) is much reduced compared to the compound optimization sets discussed above (Table 2). Data peaks are clearly distinguished from several notable activity cliffs that are also contained in the screening set. These cliffs become much more apparent in the 3D landscapes than the corresponding 2D representations shown in Figure 5b, due to the large number of data points. Figure 5c shows the structures of representative active compounds that are mapped in Figure 5a. These compounds are structurally diverse and include the most active molecules from selected compound clusters. Taken together, these results illustrate that 3D activity landscape representations are also applicable to raw screening data and clearly help to quickly focus on compound subsets that form activity cliffs and contain SAR information.

## CONCLUSIONS

Herein we have focused on generating activity landscape views for actual compound data sets that can be compared and analyzed in qualitative and quantitative terms. As we expected, details of approximated "true" activity landscapes depart from the idealized canyon/rolling hills landscape view that we utilize to rationalize principal relationships between activity landscapes and structure−activity relationships. However, we have found that different compound data sets produce different types of activity landscapes that are readily interpretable, despite molecular representation-dependent differences in their topology. Furthermore, we have found that landscape features can be related to global SAR characteristics of compound data sets deduced from systematic pairwise comparisons of compound similarity and potency and quantified by SAR discontinuity scoring. Visualizing similarity and potency relationships in three-dimensional landscape representations makes it possible to assess SAR characteristics of a compound data set and to identify activity cliffs of varying magnitude. Activity landscapes of different compound sets mirror previous findings that SARs are predominantly heterogeneous in nature and that even largely continuous SARs contain elements of discontinuity, which become apparent as shallow activity cliffs in landscape models. However, activity cliffs that occur in an activity landscape for a given molecular representation might be modified or even leveled out in a different chemical reference space. Hence, for a comprehensive description and prioritization of activity cliffs in a data set, the choice of molecular representations is rather critical. Furthermore, activity landscape visualization also provides an intuitive way to identify molecular representations that best separate highly and weakly potent molecules in a given data. Such representations are most suitable for many practical applications of molecular similarity analysis.

## ACKNOWLEDGMENT

L.P. is supported by Boehringer Ingelheim Pharma GmbH & Co. KG.

**Note Added after ASAP Publication.** This paper was published on the Web on May 5, 2010, with an error to Figure 3b. The corrected version was reposted to the Web on May 10, 2010.

**Supporting Information Available:** Figure S1 provides 2D activity landscape representations and clustering results for all seven compound data sets. This information is available free of charge via the Internet at http://pubs.acs.org.

## REFERENCES AND NOTES

(1) Maggiora, G. M. On Outliers and Activity Cliffs - Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535.
(2) Peltason, L.; Bajorath, J. Systematic Computational Analysis of Structure-Activity Relationships: Concepts, Challenges and Recent Advances. *Future Med. Chem.* **2009**, *1*, 451–466.
(3) 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.
(4) Peltason, L.; Bajorath, J. SAR Index: Quantifying the Nature of Structure-Activity Relationships. *J. Med. Chem.* **2007**, *50*, 5571–5578.
(5) Guha, R.; Van Drie, J. H. Structure-Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* **2008**, *48*, 646–658.
(6) 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.
(7) 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.
(8) 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.
(9) Wawer, M.; Peltason, L.; Bajorath, J. Elucidation of Structure-Activity Relationship Pathways in Biological Screening Data. *J. Med. Chem.* **2009**, *52*, 1075–1080.
(10) 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.
(11) Shanmugasundaram, V.; Maggiora, G. M. Characterizing Property and Activity Landscapes Using an Information-Theoretic Approach. Proceedings of 222nd American Chemical Society National Meeting, Division of Chemical Information, Chicago, IL, August 26−30, 2001; American Chemical Society: Washington, D.C., 2001; abstract no. *77*.
(12) Medina-Franco, J. L.; Maggiora, G. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. A Similarity-based Data-fusion Approach to the Visual Characterization and Comparison of Compound Databases. *Chem. Biol. Drug Des.* **2007**, *70*, 393–412.
(13) Borg, I.; Groenen, P. J. F. *Modern Multidimensional Scaling. Theory and Applications*, 2nd ed.; Springer: New York, NY, 2005.
(14) Cressie, N. *Statistics for Spatial Data*, revised ed.; Wiley: New York, NY, 1993.
(15) *MDL Drug Data Report (MDDR)*, version 2005.2; Symyx Software: San Ramon, CA, 2005.
(16) PubChem BioAssay; National Center for Biotechnology Information (NCBI): Bethesda, MD; http://pubchem.ncbi.nlm.nih.gov/ (accessed March 1, 2010).

(17) *MACCS structural keys*; Symyx Software: San Ramon, CA, 2002.
(18) *Molecular Operating Environment (MOE)*; Chemical Computing Group, Inc.: Montreal, Quebec, 2007.
(19) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naive Bayesian Classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170–178.
(20) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
(21) Gedeck, P.; Willett, P. Visual and Computational Analysis of Structure-Activity Relationships in High-Throughput Screening Data. *Curr. Opin. Chem. Biol.* **2001**, *5*, 389–395.
(22) Cooley, W.; Lohnes, P. *Multivariate Data Analysis*; Wiley: New York, NY, 1971.
(23) Oprea, T. I.; Gottfries, J. Chemography: The Art of Navigating in Chemical Space. *J. Comb. Chem.* **2001**, *3*, 157–166.
(24) Sammon, J. W. A Non-linear Mapping for Data Structure Analysis. *IEEE Trans. Comput.* **1969**, *C-18*, 401–409.
(25) Agrafiotis, D. K.; Lobanov, V. S. Nonlinear Mapping Networks. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1356–1362.
(26) Kohonen, T. *Self-Organizing Maps*; Springer: Heidelberg, Germany, 1996.
(27) Kruskal, J. B. Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. *Psychometrika* **1964**, *29*, 1–27.
(28) Kruskal, J. B. Nonmetric Multidimensional Scaling: A Numerical Method. *Psychometrika* **1964**, *29*, 115–129.
(29) Venables, W. N.; Ripley, B. D. *Modern Applied Statistics with S*, 4th ed.; Springer: New York, NY, 2002.
(30) Bates, D.; Chambers, J.; Dalgaard, P.; Falcon, S.; Gentleman, R.; Hornik, K.; Iacus, S.; Ihaka, R.; Leisch, F.; Lumley, T.; Maechler, M.; Murdoch, D.; Murrell, P.; Plummer, M.; Ripley, B.; Sarkar, D.; Temple-Lang, D.; Tierney, L.; Urbanek, S. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2008.
(31) Furrer, R.; Nychka, D.; Sain, S. *Tools for Spatial Data, R package*, version 5.01; R Foundation for Statistical Computing: Vienna, Austria, 2008.
(32) Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244.
(33) Torgerson, W. S. *Theory and Methods of Scaling*; Wiley: New York, NY, 1958.

CI100091E

# Summary

Herein, a novel computational approach for the generation of 3D activity landscapes has been reported. In addition, these 3D activity landscapes generated for different compound sets were analyzed in detail. Such *true* landscapes displayed substantial variations compared to conceptualized smooth/rugged views frequently employed to explain SAR features. Moreover, the generated 3D landscapes showed notable dependence on the chosen molecular representation and variation in their topologies. In order to enable direct comparisons, the landscapes were represented within a consistent framework. Despite clear differences, the overall SAR characteristics obtained after systematic pairwise comparison of structural similarities and activity differences, correlated well with the landscape features. In addition, the activity cliffs were readily identifiable, although, they were dependent on the chosen molecular representation. The 3D activity landscapes also aided in the intuitive identification of chemical reference spaces that allowed the separation of compounds according to their potency levels. Therefore, these landscape representations were well suited for qualitative analysis of SAR.

Following the successful generation of 3D activity landscape representations for real data sets, a follow-up study was performed to compare these with standard 2D representations reported in the next chapter.

# Chapter 2

# Comparison of two- and three-dimensional activity landscape representations for different compound data sets

## Introduction

The SARs associated with bioactive compound data are routinely investigated with the help of the activity landscape concept as it is intuitive and fairly easy to interpret. The important advantage provided by the landscape representations is the relative ease in the visual accessibility of characteristic SAR features prevalent in various sets of active compounds. Indeed, several methodologies for activity landscape modeling have been developed. The common objective of these conceptually different modeling approaches is to combine the pairwise chemical similarity and activity relationships existing in a given data set. Herein, a comparative study has been outlined where exemplary 2D activity landscape representations, i.e. NSGs, generated for data sets with different SAR phenotypes were compared with the 3D models. The study clearly revealed that both 2D and 3D landscapes capture the overall SAR content of the data sets used in an analogous manner, despite their distinctive topologies. Additionally, it has also be observed that local SAR features are perceived differently in these representations.

# Comparison of two- and three-dimensional activity landscape representations for different compound data sets

**Preeti Iyer, Mathias Wawer and Jürgen Bajorath***

***Received 25th October 2010, Accepted 15th November 2010***
**DOI: 10.1039/c0md00188k**



Modeling of activity landscapes provides a basis for the analysis of structure–activity relationships (SARs) in large compound data sets. Activity landscape models enable visual access to SAR features. Regardless of their specific details, these models generally have in common that they integrate molecular similarity and potency relationships between active compounds. Different two-dimensional (2D) landscape representations have been introduced and recently also the first detailed three-dimensional (3D) model. Herein we compare advanced 2D and 3D activity landscape models for compound data sets having different SAR character. Although the compared 2D and 3D representations are conceptually distinct, it is found that global SAR features of compound data sets can be equally well deduced from them. However, local SAR information is often captured in different ways by these representations. Since these 2D and 3D landscape modeling tools have been made freely available, the analysis also provides guidelines for how to best utilize these alternative landscape representations for practical SAR analysis.

## Introduction

The study of compound structure–activity relationships plays a central role in medicinal chemistry, and a variety of computational approaches are employed to analyze and predict SARs in a qualitative or quantitative manner.[1–9] Among these approaches is the modeling of *activity landscapes* for compound data sets.[10,11] Generally, an activity landscape can be understood as any representation that integrates similarity and potency relationships between compounds sharing the same biological activity.[11] Different types of 2D activity landscape representations have been introduced.[6,11] In addition, 3D landscape models have also been discussed[6,11] that can be rationalized as a 2D projection of a chemical reference space (where proximity of compounds is an indicator of similarity) with compound potency added as the third dimension. Such 3D landscape views then essentially describe biological activity response surfaces as a consequence of changes in chemical structure (*i.e.* "walks" in chemical space[12]).

The most prominent feature of activity landscapes, however they might be represented, are *activity cliffs*[10,11,13] that are formed by structurally similar compounds or analogs with large differences in potency. Activity cliff regions are associated with high SAR information content[11,13,14] because in these regions small changes in compound structure lead to significant potency effects.

Although several 2D activity landscape representations of varying complexity are currently available, only hypothetical 3D activity landscapes have been discussed until recently when the first detailed 3D landscape model has been introduced for the analysis of compound data sets.[15] Alternative activity landscape representations have thus far rarely been compared for given compound data sets to better understand how they capture SAR information in detail and how they might differ.

Therefore, we have generated conceptually distinct 2D and 3D activity landscape views of different compound data sets, and using different molecular representations, in order to compare how these models convey SAR information. The results presented herein show that distinct landscape models often capture SAR information contained in compound data sets in a comparable manner, but also reveal unique features of 2D and 3D landscape representations and their SAR information content. Furthermore, our analysis also provides some practical guidelines for the complementary use of different activity landscape models in medicinal chemistry.

## Methods

### 3D activity landscapes

The generation of 3D activity landscapes followed a previously reported protocol.[15] Euclidean distance relationships between compounds (constituting a coordinate-free chemical reference space) were calculated using the MACCS[16] or, alternatively, Molprint2D[17] fingerprint and projected onto a 2D plane using multidimensional scaling (MDS)[18] as a dimension reduction technique. Compound coordinates were normalized to range from 0 to the maximum observed pairwise chemical dissimilarity in a given compound data set, such that the range of the planar coordinates (and hence the size of the landscape plots) reflected the overall chemical dissimilarity within the data set. Compound potency information was then added as a third dimension and potency values were interpolated using the Krige function[19] to create coherent surfaces. To simplify the comparison of different landscapes, all representations were displayed from the same view point (azimuth: 60°, co-latitude: 45°). The potency axis was

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

consistently scaled for all data sets, ranging from the lowest (3.46) to the highest (11.72) log-potency value observed for all compound activity classes. Furthermore, *x*- and *y*-axes ranged from the lowest (0.00) to the highest values of pairwise chemical distances of all activity classes (MACCS: 9.27, Molprint2D: 9.79). The surface of the activity landscapes was colored according to interpolated potency values (surface elevation) using a color gradient from green (lowest potency) to red (highest potency). The value range mapped onto the color gradient was determined by the highest minimal and the lowest maximal interpolated potency values of all activity classes, ranging from a log-potency of 5.79 (green) to 9.19 (red). In addition, landscape transparency[15] was adjusted to reflect the density of experimental potency measurements. Grid points with close proximity to original data points (compounds) were colored opaque and those furthest away from data point were fully transparent. Hence, regions of interpolated surface area not populated with data points appear in white.

### Network-like similarity graphs

A Network-like similarity graph (NSG)[20] is a 2D activity landscape model that represents similarity and potency relationships in a compound data set as an annotated graph. In NSGs, nodes represent individual molecules that are connected by edges if their structural similarity exceeds a predefined threshold. Here, dissimilarity relationships corresponding to similarity threshold criteria were also expressed by Euclidean distances to enable a direct comparison with 3D landscape models. Euclidean distance thresholds of 4.90 and 5.29 were used for MACCS and Molprint2D, respectively (corresponding to pairwise MACCS Tanimoto similarity[21] of 0.65), which also yielded a comparable edge density in all NSGs. Nodes were colored by potency corresponding to the color scheme of 3D landscapes based on a continuous gradient from green (log-potency $\leq 5.57$) *via* yellow to red (log-potency $\geq 10.05$). The lower and upper boundary of the gradient corresponded to the highest minimal and the lowest maximal potency values of all activity classes. Nodes were scaled in size based on a numerical SAR analysis function, the local SAR Index (SARI),[4,20] which determines the contribution of an individual compound to global SAR discontinuity. Generally, SAR discontinuity is introduced when small changes in compound structure are accompanied by large changes in potency. Thus, in NSGs, large nodes indicate that the potency of a compound substantially differs from that of its structural neighbors. For NSG display, a graphical layout algorithm is applied that places multiple densely connected compounds in close vicinity and separates weakly connected regions from each other.[22]

### Compound activity classes

For activity landscape comparison, five data sets consisting of 112–209 compounds active against different targets were taken from the MDDR,[23] as summarized in Table 1. These data sets were selected based on global SARI[4] calculations to represent different degrees of SAR heterogeneity.

**Table 1** Compound activity classes[a]

| Abbreviation | Activity class | No. of compounds | Potency range/nM |
|---|---|---|---|
| ACH | Acetylcholinesterase | 112 | 0.02–85000 |
| COX | Cyclooxygenase 2 | 149 | 0.09–50000 |
| 5HT | 5-HT reuptake | 129 | 0.01–2700 |
| PH4 | Phosphodiesterase IV | 209 | 0.0025–348000 |
| THR | Thrombin | 172 | 0.0019–30000 |

[a] For the five activity classes used in the study, the number of compounds and their potency range are reported.

## Results and discussion

### Two-dimensional activity landscape representations

In Fig. 1, alternative 2D landscape views are shown that illustrate their conceptual diversity. All these representations have in common that they integrate structure and potency relationships that are present in compound data sets, albeit in different ways. Structure–activity similarity (SAS) maps[10] have probably been the first explicit 2D activity landscape representation. SAS maps compare structure and activity similarity of active compounds in a pairwise manner. Activity similarity is expressed by potency differences. These maps delineate regions of high SAR discontinuity (*i.e.* high structural and low activity similarity) and high SAR continuity (*i.e.* low structural and high activity similarity) as well as nondescript regions that contain only very little SAR information (*e.g.* where both structure and activity similarity are low). Also shown is a color coded 2D projection of chemical reference space (providing the basis for 3D landscape modeling; see Methods) that captures compound distance relationships and mirrors potency distributions. Thus, the information provided is similar to SAS maps, but the representations are distinctly different. In addition, a so-called structure–activity landscape index (SALI) graph[5] is shown that follows a different design idea and represents a potency-directed compound network. SALI graphs largely focus on identifying activity cliffs of different magnitude and delineating cliff pathways for different SALI score threshold values. Furthermore, in network-like similarity graphs,[20] compounds are also represented as nodes and undirected edges indicate pairwise similarity relationships. In NSGs, color coding provides potency information and scaling of nodes the contribution of each individual compound to local SAR discontinuity (*i.e.* the larger the node, the higher the degree of discontinuity a compound introduces). Thus, prominent activity cliffs are indicated by connected pairs of large red and green nodes. Different from SALI graphs and other representations, NSGs are designed to explore both global and local SAR features in compound data sets and identify compound subsets that form different SARs patterns not limited to activity cliffs. As such, NSGs probably represent the most detailed 2D activity landscape views that are available at present.

### Three-dimensional activity landscapes

In Fig. 2, an idealized 3D activity landscape and a 3D landscape calculated for a specific compound data set are shown. Idealized activity landscapes have frequently been utilized to illustrate

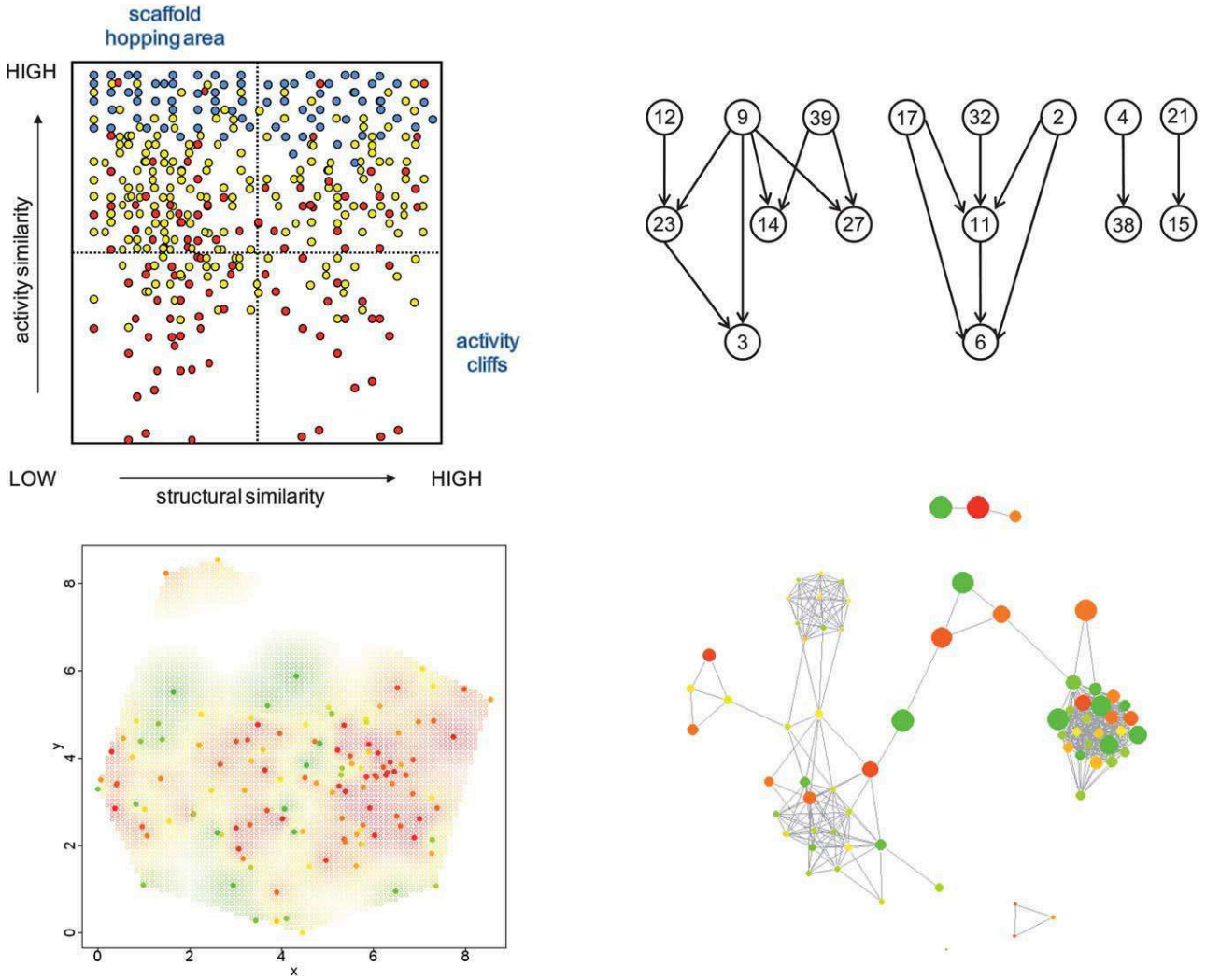



**Fig. 1** Alternative 2D activity landscape representations. Four conceptually different graphical representations for SAR analysis are shown. SAS maps (upper left) represent all pairs of compounds of a data set as points in a scatter plot organized by the level of structural and activity similarity for each compound pair. The color of points reflects whether the potency level of the more potent compound of the pair is high (blue), intermediate (yellow), or low (red). In all other representations, points or nodes represent individual molecules. In SALI graphs (upper right), the nodes are connected by edges if they form an activity cliff exceeding a predefined SALI score. The plot in the lower left section was generated by mapping compound dissimilarity values to 2D coordinates using multi-dimensional scaling. The gradient from green to red reflects potency from low to high values. The same color code is applied for the nodes in an NSG (lower right). Here, edges between nodes are drawn if the similarity of the corresponding compounds exceeds a predefined threshold. In addition, nodes are scaled in size according to local SAR discontinuity contributions of the corresponding compounds.

important SAR characteristics including activity cliffs as well as rugged regions and smooth regions that represent SAR discontinuity and continuity, respectively. Hence, in such 3D landscape views, differences in landscape topology can be readily associated with different SAR behaviour. Despite the intuitive nature of 3D activity landscapes, 3D landscape models for actual compound data sets have only recently been introduced.[15] The exemplary data set-based activity landscape shown in Fig. 2 mirrors characteristic topological features depicted in idealized landscapes.

### Comparison of NSGs and 3D landscape models

Despite their different design, the 2D and 3D landscape models presented herein have common features. Both NSGs and 3D landscapes rely on the assessment of global molecular similarity (other similarity measures would of course also be possible). Furthermore, both 2D and 3D landscape models capture global SAR features and local SAR environments. We have compared these representations in detail for five different compound activity classes summarized in Table 1. In each case, activity landscapes were generated with two different molecular representations; MACCS keys, a structural fragment fingerprint, and Molprint2D, a combinatorial fingerprint that captures layered topological atom environments. For each compound data set, the two pairs of corresponding NSGs and 3D models are shown in Fig. 3.

The general observation can be made that both 2D and 3D activity landscape features were in part significantly influenced

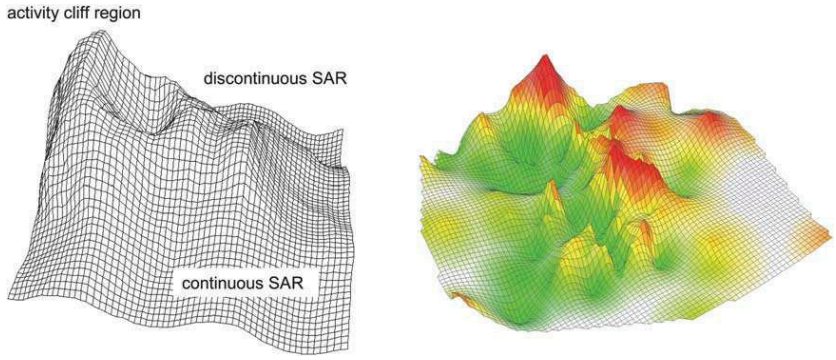



**Fig. 2** Hypothetical and compound data-based 3D activity landscapes. The landscape on the left represents an idealized heterogeneous activity landscape that contains different local SAR regions. High peaks correspond to activity cliffs, rugged regions represent discontinuous local SARs (small changes in structure are accompanied by significant potency effects) and smooth regions continuous local SARs (gradual changes in structure are accompanied by small changes in potency). On the right, a 3D landscape model calculated on the basis of an actual compound data set (acetylcholinesterase inhibitors in Table 1) is shown. The surface of the activity landscapes is colored according to interpolated potency values (surface elevation) using a color spectrum from green to red.

by the chosen fingerprint representation, in accord with earlier findings.[15] For all compound classes, activity cliffs produced by the alternative molecular representations often differed, due to the fact that the two fingerprints often accounted for compound similarity relationships in different ways. In this context, it should be noted that the topology of NSGs is only determined by pairwise similarity relationships and the graphical layout algorithm that separates densely connected compound clusters for visualization. Hence, in NSGs, fingerprint-specific differences in the distribution of similarity values were rather obvious.

Another general observation has been that NSGs and 3D landscape models provided essentially equivalent global views of global SAR characteristics of different data sets, despite representation-dependent differences. For example, in Fig. 3a, the NSGs and 3D landscapes of class ACH (the MACCS-based 3D model of this data set is also shown in Fig. 2) both display coexisting continuous and discontinuous regions interspersed with activity cliffs. The COX landscapes in Fig. 3b are characterized by the presence of many structurally similar compounds with relatively low potency and a higher degree of global continuity. By contrast, the 5HT landscapes in Fig. 3c are much more discontinuous in nature (here, fingerprint-dependent topology differences are striking) and contain much larger activity cliffs. Moreover, all PH4 landscapes in Fig. 3d are characterized by the presence of many activity cliffs, which represent their most significant feature. Similarly, the THR landscapes in Fig. 3e are also strongly discontinuous in nature. In this case, MACCS (but not Molprint2D) introduced a notable compound clustering effect that can be well appreciated in both the NSG and the corresponding 3D model. The NSG displays a clear separation of densely connected clusters and the 3D landscape contains a large area of purely interpolated white surface separating two compound subsets.

Taken together, the comparisons shown in Fig. 3 revealed that differences in SAR information content between analyzed compound data sets and their SAR characteristics were well accounted for by both 2D and 3D activity landscape representations. Moreover, these landscape views could also be used as a diagnostic for chosen molecular representations. For example, in the case of THR, comparison of the MACCS- and Molprint2D-based landscapes clearly indicated that the choice of one or the other fingerprint would lead to substantial differences in the analysis of structure–activity relationships.

## Mapping of activity cliffs

We next focused our analysis on the comparison of activity cliffs in our 2D and 3D landscape models. NSGs are based on calculated pairwise compound distances and measured potency values, whereas 3D activity landscapes are based on projected compound distances and interpolated potency values. Hence, the data structure underlying 3D landscapes is in principle more approximate in nature than the NSG data structure. However, the topology of 3D landscapes provides a particularly intuitive access to prominent activity cliffs and we therefore selected large cliffs from these 3D landscape views and then mapped these cliffs to NSGs. The results are also shown in Fig. 3. With no exception, prominent activity cliffs selected from 3D activity landscape models were also found to be large activity cliffs in NSGs. Thus, corresponding activity cliffs in 3D models and NSGs were of comparably large magnitude. For activity classes with moderate or low global SAR discontinuity (ACH and COX) the largest activity cliffs were easily identified in both 2D and 3D representations. For the remaining classes with increasingly high discontinuity, many comparably large activity cliffs were observed in both 2D and 3D landscape representations.

## Differences in SAR information content

The comparisons in Fig. 3 also reveal a number of differences in the way NSGs and 3D activity landscapes capture SAR information. For example, similarity relationships between individual compounds and their local SAR contributions are only provided by NSGs. Furthermore, NSG topology is ultimately determined by edge connectivity, whereas 3D landscapes are based on explicit pairwise compound distances. This difference generally results in stronger compound clustering effects observed in NSGs. Furthermore, the topology of NSGs is not affected by "structural outliers" that do not form similarity relationships to other compounds, whereas such outliers might influence the

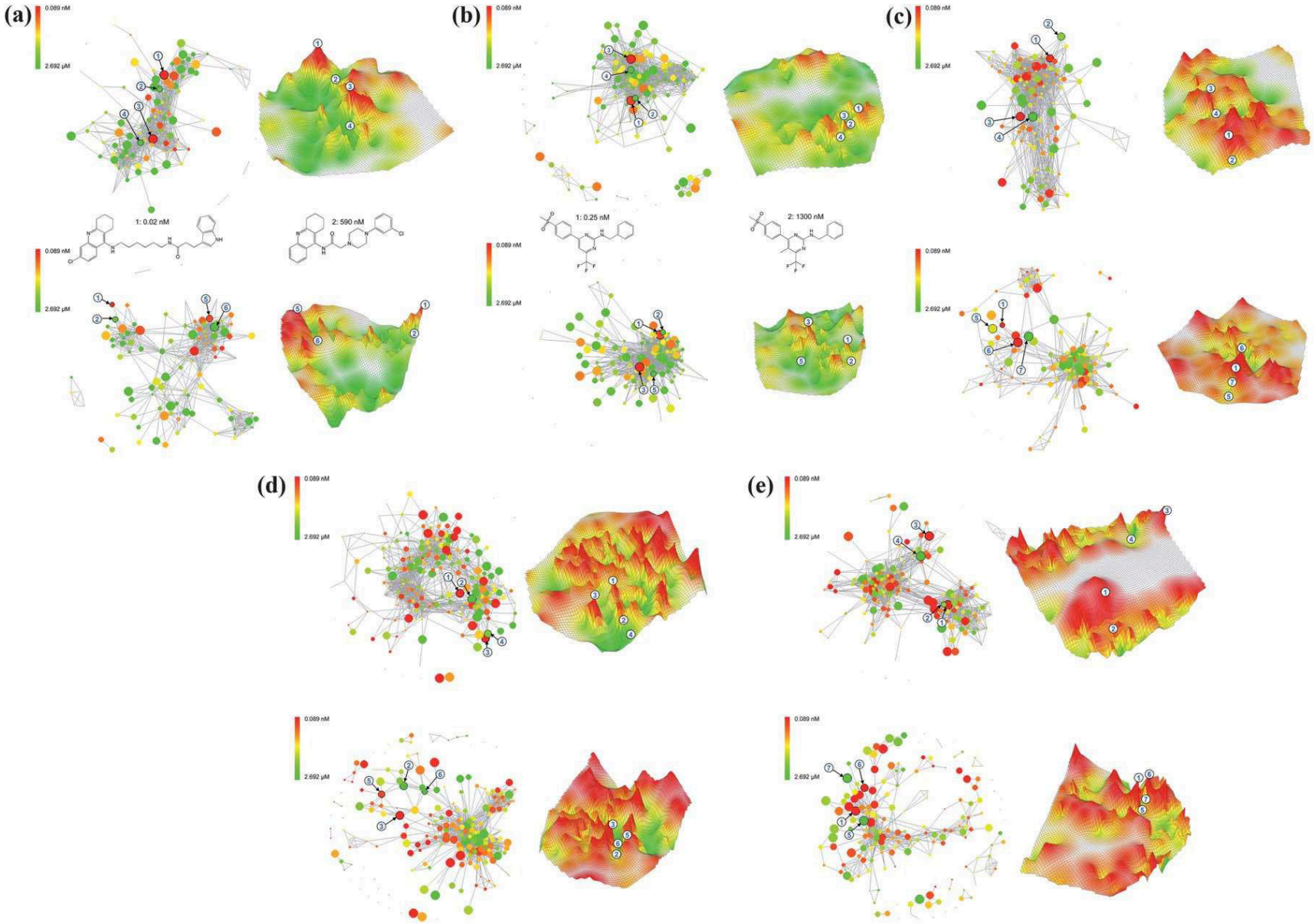



**Fig. 3** Comparison of NSGs and 3D activity landscape models. NSG and 3D landscape representations calculated on the basis of either MACCS (top) or Molprint2D fingerprint distances (bottom) are presented for five different compound data sets. Corresponding positions of selected compounds in 2D and 3D representations are indicated by numbers. Exemplary compound structures are also displayed. Landscape representations are shown for sets of (a) acetylcholinesterase inhibitors (ACH), (b) cyclooxygenase 2 inhibitors (COX), (c) serotonin reuptake inhibitors (5HT), (d) phosphodiesterase 4 inhibitors (PH4), and (e) thrombin inhibitors (THR).

topology of 3D models. By contrast, areas of sparse SAR data are much easier to identify in 3D landscape models as purely interpolated surface area, which provides a basis for further directed compound data collection. Moreover, 3D landscape models are better suited to quickly focus on the most prominent activity cliffs in a data set than NSGs, in particular, in the presence of strong global SAR discontinuity. This is the case because both surface elevation and the color gradient of 3D landscape models mark prominent activity cliffs. However, the shape of large-magnitude activity cliff regions is the most difficult surface area to interpolate and hence 3D landscapes contain little interpretable information concerning the immediate environment of activity cliffs. This information is provided in detail in NSGs, which suggests the complementary use of 3D activity landscape models and NSGs for the detailed analysis of activity cliff regions in compound data sets. Because the intrinsically different layout of NSGs and 3D landscape models, it is usually difficult to delineate corresponding regions (compound subsets) in these representations. However, once prominent activity cliffs have been identified in 3D landscape models, they can directly be mapped to NSGs where their local SAR environments can then be analyzed in detail and from which other attractive candidate compounds can be selected.

## Conclusions

Activity landscape models are attractive tools for graphical SAR analysis of large compound data sets. Once generated, such graphical representations can be easily navigated and interpreted by medicinal chemists. Activity landscapes can be modeled in two or three dimensions. However, while idealized 3D landscape models have already been used for considerable time to rationalize SAR characteristics, only recently detailed 3D activity landscape models have been derived for sets of known active compounds. These models are obtained by dimension reduction of computational chemical reference spaces followed by interpolation of potency surfaces, and their most attractive feature for visual analysis is their intuitive topology. Herein, we have compared in detail 3D landscape models for different data sets with corresponding network-like similarity graphs, which provide a distinctly different access to SAR data. For model comparison, we have established a consistent data representation scheme. Global SAR characteristics and differences between data sets could be well appreciated on the basis of both 2D and 3D landscape models. Also, especially the 3D landscape models clearly showed how much SAR analysis might depend on the chosen molecular representations (here different types of fingerprints). Accordingly, these models can be used as a diagnostic tool to better understand how to represent data sets for meaningful SAR analysis. For example, for a systematic exploration of SARs in large data sets, representations are preferred that produce contiguous landscape surfaces and do not result in strong compound clustering effects. Furthermore, both 2D and 3D landscape models could be applied to monitor evolving compound data sets and gain insights into the progression of SAR trends. Moreover, especially 3D landscape models might also be utilized to identify under-sampled SAR regions in data sets.

Importantly, we found that large-magnitude activity cliffs identified in 3D landscape models consistently corresponded to large activity cliffs in NSGs. For a detailed analysis of activity cliffs, 3D and 2D landscapes are best used in a complementary manner because prominent activity cliffs are easily identified in 3D models and details of their SAR environments can then be extracted from NSGs.

The NSG tools are publicly available as part of the SARANEA software[24] and programs to generate 3D activity landscape models are also freely available (both can be obtained *via* http://www.lifescienceinformatics.uni-bonn.de/; see the Downloads section). The comparative activity landscape analysis presented herein should be helpful to further study specific features of 2D and 3D landscapes and also provides some guidelines how to utilize these landscape representations for practical SAR analysis.

## References


1 D. T. Manallack, D. D. Ellis and D. J. Livingstone, *J. Med. Chem.*, 1994, **37**, 3758–3767.
2 D. E. Patterson, R. D. Cramer, A. M. Ferguson, R. D. Clark and L. E. Weinberger, *J. Med. Chem.*, 1996, **39**, 3049–3059.
3 E. X. Esposito, A. J. Hopfinger and J. D. Madura, *Methods Mol. Biol.*, 2004, **275**, 131–214.
4 L. Peltason and J. Bajorath, *J. Med. Chem.*, 2007, **50**, 5571–5578.
5 R. Guha and J. H. Van Drie, *J. Chem. Inf. Model.*, 2008, **48**, 646–658.
6 L. Peltason and J. Bajorath, *Future Med. Chem.*, 2009, **1**, 451–466.
7 A. R. Leach, V. J. Gillet, R. A. Lewis and R. Taylor, *J. Med. Chem.*, 2010, **53**, 539–558.
8 H. Geppert, M. Vogt and J. Bajorath, *J. Chem. Inf. Model.*, 2010, **50**, 205–216.
9 M. Wawer and J. Bajorath, *J. Chem. Inf. Model.*, 2010, **50**, 1395–1409.
10 V. Shanmugasundaram and G. M. Maggiora, Proceedings of 222nd American Chemical Society National Meeting, *Division of Chemical Information*, 2001; abstract no. 77.
11 A. M. Wassermann, M. Wawer and J. Bajorath, *J. Med. Chem.*, in press, DOI: 10.1021/jm100933w.
12 R. van Deursen and J.-L. Reymond, *ChemMedChem*, 2007, **2**, 636–640.
13 G. M. Maggiora, *J. Chem. Inf. Model.*, 2006, **46**, 1535–1535.
14 G. M. Maggiora and V. Shanmugasundaram, *Methods Mol. Biol.*, 2011, **672**, 39–100.
15 L. Peltason, P. Iyer and J. Bajorath, *J. Chem. Inf. Model.*, 2010, **50**, 1021–1033.
16 *MACCS Structural Keys*; Symyx Software: San Ramon, CA, 2005.
17 A. Bender, H. Y. Mussa, R. C. Glen and S. Reiling, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 170–178.
18 I. Borg and P. J. F. Groenen, Modern Multidimensional Scaling. *Theory and Applications*, 2nd ed.; Springer: New York, NY, 2005.
19 N. Cressie, *Statistics for Spatial Data*, revised ed.; Wiley: New York, NY, 1993.
20 M. Wawer, L. Peltason, N. Weskamp, A. Teckentrup and J. Bajorath, *J. Med. Chem.*, 2008, **51**, 6075–6084.
21 P. Willett, J. M. Barnard and G. M. Downs, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 983–996.
22 T. M. J. Fruchterman and E. M. Reingold, *Software: Pract. Exper.*, 1991, **21**, 1129–1164.
23 *MDL Drug Data Report (MDDR)*; Symyx Software: San Ramon, CA, 2005.
24 E. Lounkine, M. Wawer, A. M. Wassermann and Bajorath, *J. Chem. Inf. Model.*, 2010, **50**, 68–78.

## Summary

A comparative analysis performed using the activity landscape representations generated by two conceptually different computational techniques has been reported. The 2D representations, i.e. the NSGs, integrate systematic similarity and potency relationships in data sets in the form of a network and therefore, do not require dimension reduction. On the other hand, 3D activity landscapes are contiguous surfaces generated after interpolation of activity information of constituent compounds and their 2D projection obtained as a result of dimension reduction. From the study, it was observed that in spite of their visual differences, these representations perceived global SAR content in a similar manner. In addition, 3D models demonstrated the influence exerted by the choice of the molecular presentations on the underlying chemical spaces and as a result the analysis of SAR. The 2D and 3D models provided different perspectives during the analysis of local SAR environments, as expected. Activity cliffs of large magnitude that represent centers of SAR discontinuity information were readily identifiable using 3D representations. However, SAR exploration in the vicinity of such cliffs was better facilitated by NSGs. Prominent activity cliffs identified in NSGs were also found to be consistently represented in the 3D models. Thus, this study clearly indicated that the complementarity of these representations can be exploited during detailed exploration of activity cliffs. It is important to note that the comparison of different activity landscape representations provides valuable indicators as to how these models can be maximally utilized during practical SAR investigations.

Pairwise comparisons of compound structures and activities form the core aspect of activity landscape models despite differences in their conceptual design. Accordingly, various landscape features (e.g. activity cliffs) are typically characterized at the level of compound pairs. A computational approach that resolves different landscape features at the level of individual compounds has been introduced in the following chapter.

# Chapter 3

# Conditional probabilities of activity landscape features for individual compounds

## Introduction

Activity landscape modeling is frequently utilized during SAR analysis of large data sets as it combines pairwise chemical similarities and potency differences between active compounds [1]. Various computational methodologies for the construction of landscape representations to study single- as well as multi-target SAR have been introduced, ranging from molecular networks and 2D plots to 3D landscape models [1, 2]. In spite of their conceptual differences, these approaches for activity landscape generation require systematic pairwise comparisons of molecular similarities and potency differences. Therefore, the resulting landscape features are often resolved at the level of compound pairs. Activity cliffs, for instance, are formed by structurally related compound pairs with large variation in their potencies and represent the most significant landscape features [3, 4]. Additionally, structurally similar and dissimilar compound pairs having equivalent potencies also constitute other landscape features relevant for SAR analysis.

Various modeling techniques provide a qualitative description of activity landscape features. In a recent study, these features have quantitatively been

characterized using an information-theoretic approach [5]. Information entropy calculations were performed to characterize the information associated with different landscape features and the results were compared with their SAR information content.

On the basis of probability theory, a methodology has recently been introduced for the quantitative assessment of predefined activity landscape features [6]. Conditional probabilities for individual compounds to form these landscape features were computed using their pairwise structural similarity and potency difference data. This computational approach was applied to several data sets and compounds with significant feature probabilities were identified. In addition, compounds were assigned to various feature categories on the basis of their conditional probabilities. The conceptual advantages of conditional probability calculations and the resulting compound assignments have been highlighted with the help of graphical landscape representations [6]. The study reported herein has been published in reference [6] of this chapter. My contributions to this study included the systematic calculation of conditional feature probabilities for different data sets as well as the generation of exemplary graphical representations.

# Methodology

## Similarity-Activity Similarity Maps

The activity landscape features have initially been categorized with the help of SAS maps, the first 2D landscape representations [7]. SAS maps represent activity landscapes at the level of compound pairs and are appropriately suited for the classification of landscape features. SAS maps were originally implemented as 2D plots of structural versus activity similarity of molecules within a data set obtained from systematic pairwise comparisons.

Typically, chemical relatedness is determined by comparing molecular fingerprints using the Tanimoto coefficient (Tc) [8] as the similarity metric and activity similarity is expressed in terms of potency differences (normalized, if required) between all compound pairs. Structural similarity is represented along the x-axis while the y-axis reports the absolute differences in activity. Different

SAS map variants have been proposed for various practical applications [9, 10]. Data points in the SAS map represent similarity and potency relationships between compound pairs. The map can be divided into four distinct regions on the basis of predefined chemical similarity and activity difference thresholds.

## Similarity and Activity Difference Thresholds

Clear definition of threshold values for structural similarity and potency difference is a necessary requirement for the quantitative analysis of activity landscape features [1, 4]. The chemical similarity was assessed using extended connectivity fingerprints [11] with a bond diameter 4 (ECFP4) and a Tc value of 0.55 served as the similarity threshold while absolute potency difference threshold was set to 2 pKi units corresponding to 2 orders of magnitude (OoM) difference in activity [6]. These threshold values have often been used to define activity cliffs in various computational studies [4, 12]. Compound pairs with ECFP4 Tc values of 0.55 or higher are typically structurally related [13] and absolute potency differences of 2 OoM or more account for significant activity cliffs within a data set [12].

## Activity Landscape Features

The predefined molecular similarity and potency difference thresholds divide the SAS maps into four distinct sections that represent the different activity landscape features. The lower left portion is composed of compound pairs that are chemically dissimilar but have similar potencies. Such pairs consist of compounds having different molecular scaffolds or otherwise dissimilar structures and are referred to as *similarity cliffs* [5]. Therefore, molecule pairs forming similarity cliffs are located far apart in the chemical reference space.

Pairs of compounds contained in the lower right region exhibit high chemical similarity and low potency differences. These are typically associated with smooth landscape areas and accordingly, known as *smooth pairs*. The upper right part of the SAS map comprises of *activity cliffs*, i.e., structurally related compound pairs having large potency differences. The entire right section of the SAS map is composed of molecule pairs that are located in close proximity in the chemical space due to their high structural similarities. The upper left section

containing structurally dissimilar compound pairs with large potency differences forms the nondescript region and has relatively low SAR information. Hence, *similarity cliffs*, *smooth pairs* and *activity cliffs* were considered as principal activity landscape features [6].

Characterization of the activity landscape features derived from SAS maps using compound pair frequencies and information entropy calculations has been performed in a previous study [5]. From a statistical perspective, activity cliffs are considered most informative due to their sparse distribution. By contrast, similarity cliffs have relatively low information as they occur with high frequency. However, from an SAR point of view, similarity cliffs provide important information about chemically dissimilar compounds with similar potencies. Smooth pairs are observed less frequently than similarity cliffs and therefore, have moderate information content.

## Per-compound Feature Probabilities

The core objective of the approach described in [6] has been the derivation of landscape feature probabilities for individual compounds. *The propensities of every compound in a data set to form activity cliffs, similarity cliffs or smooth pairs have been determined as feature probabilities. Feature combinations have also been considered.* Compound-based or *local* SAS maps that report pairwise similarity and potency relationships formed by specific molecules have been introduced in support of this methodology. Thus, using the information present in a local SAS map for a given molecule, the frequencies with which it participates in the formation of activity cliffs, similarity cliffs or smooth pairs have been determined [6].

## Crisp and Fuzzy Boundaries

Using predefined thresholds, pairwise structural similarity and potency difference relationships formed by a particular compound can be assigned to any one of the four different regions according to their values. Nevertheless, assignments of compound pairs to specific map regions are prone to boundary effects if precise bounds as mentioned previously are applied. For example, small variation in similarity might determine if a compound pair falls into activity cliff or

nondescript region. Furthermore, minute changes in potency differences might distinguish between pairs of compounds classified as activity cliffs or smooth pairs. Such dramatic shifts in classification resulting from small transitions about the thresholds are not chemically meaningful.

Therefore, *crisp* threshold values have been replaced with *fuzzy* boundaries, in order to balance the boundary effects that negatively affect feature assignments. *Twilight zones* or boundary intervals have been introduced in order to assign a weighted joint membership for neighboring landscape regions to the compound pairs falling into these areas [6]. Due to its mathematical foundation in fuzzy set theory concept [14], this approach made it possible to adhere to the original data partitioning scheme of local SAS maps while softening the boundaries between the different regions.

Thus, a Tc range of 0.45 to 0.65 represented the boundary interval for structural similarity while a range of one to two OoM marked the twilight region for potency difference. The weighting schemes for compound pairs within the similarity and potency difference twilight zones defined *partial memberships* to neighboring regions and produced values that ranged between 0 and 1. Accordingly, the fractional frequencies were also obtained for compounds when the magnitude of partial memberships within the twilight regions was less than one [6].

## Conditional Probabilities for Fuzzy Landscape Features

An additional conditioning has been applied so that the ability of a given compound to form similarity cliffs or activity cliffs is evaluated only with respect to other compounds having similar potencies or similar structures, respectively. The resulting conditional probabilities can be calculated from the probability of a given feature by relating it to the frequencies of all compound pairs satisfying the conditional relationship.

Hence, the respective conditional probabilities for a given compound $k$ participating in a set $V_k$ of compound pairs have been generated as follows:

1. For a pair of compounds having similar potencies,
   a. the probability to form a similarity cliff, $P(\tilde{S}^{<}|\tilde{A}^{<}, V_k)$, is given by:

$$P(\tilde{S}^{<}|\tilde{A}^{<}, V_k) = \frac{|\tilde{R}00(k)|}{|\tilde{R}00(k)| + |\tilde{R}10(k)|}$$

   b. the probability to form a smooth pair, $P(\tilde{S}^{\geq}|\tilde{A}^{<}, V_k)$, is given by:

$$P(\tilde{S}^{\geq}|\tilde{A}^{<}, V_k) = \frac{|\tilde{R}10(k)|}{|\tilde{R}00(k)| + |\tilde{R}10(k)|}$$

2. and for a structurally similar compound pair,
   a. the probability to form a smooth pair, $P(\tilde{A}^{<}|\tilde{S}^{\geq}, V_k)$, is given by:

$$P(\tilde{A}^{<}|\tilde{S}^{\geq}, V_k) = \frac{|\tilde{R}10(k)|}{|\tilde{R}10(k)| + |\tilde{R}11(k)|}$$

   b. the probability to form an activity cliff, $P(\tilde{A}^{\geq}|\tilde{S}^{\geq}, V_k)$, is given by:

$$P(\tilde{A}^{\geq}|\tilde{S}^{\geq}, V_k) = \frac{|\tilde{R}11(k)|}{|\tilde{R}10(k)| + |\tilde{R}11(k)|}$$

where $|\tilde{R}00(k)|$, $|\tilde{R}10(k)|$ and $|\tilde{R}11(k)|$ correspond to the feature probabilities for the formation of similarity cliffs, smooth pairs and activity cliffs, respectively [6].

Although the conditional probabilities given by equations *1b* and *2a* estimate the ability of a compound to form smooth pairs, they are distinct because the probability in the first case is calculated for all compound pairs with similar potency differences while in the second, it is computed for all structurally similar pairs of compounds [6].

If the denominators of the conditional probabilities specified above become very small, artificially high probabilities might be obtained. For example, this situation would apply to the conditional probabilities calculated using equa-

tions *2a* and *2b* if a compound had very few structural neighbors. Thus, denominators ($|\tilde{R}00(k)|+|\tilde{R}10(k)|$ and $|\tilde{R}10(k)|+|\tilde{R}11(k)|$) less than 2.0 were not considered for probability calculations. In order to identify significant probabilities, thresholds $P_T(\tilde{S}^{<}|\tilde{A}^{<}, V_k)$, $P_T(\tilde{S}^{\geq}|\tilde{A}^{<}, V_k)$, $P_T(\tilde{A}^{<}|\tilde{S}^{\geq}, V_k)$, $P_T(\tilde{A}^{\geq}|\tilde{S}^{\geq}, V_k)$, were determined at the $90^{\text{th}}$ percentile using sorted conditional probabilities $P(\tilde{S}^{<}|\tilde{A}^{<}, V_k)$, $P(\tilde{S}^{\geq}|\tilde{A}^{<}, V_k)$, $P(\tilde{A}^{<}|\tilde{S}^{\geq}, V_k)$, $P(\tilde{A}^{\geq}|\tilde{S}^{\geq}, V_k)$, respectively for a representative collection of data sets. Probabilities greater than their corresponding threshold values were considered significant [6].

Assessment of the respective significant conditional probabilities using a large collection of data sets allows for the identification of exceptional probabilities without taking into account their absolute magnitudes. For instance, due to the relatively rare occurrences of activity cliffs, only very few compounds in a data set are expected to form activity cliffs with their structurally similar neighbors. As a result, comparatively low probabilities might be significant for activity cliffs, although these probabilities would be substantially lower than those for similarity cliffs that typically dominate activity landscapes.

Furthermore, the use of conditional feature probabilities is also conceptually advantageous, especially for data sets of small size. For example, the conditional probability of a specific compound to form activity cliffs only takes structural neighbors into consideration and the calculation is independent of the dissimilarity relationships formed by this compound. The absolute feature probabilities on the other hand are greatly affected by the number of dissimilar compounds, making them difficult to interpret. These also reflect poorly on the ability of a compound to form activity cliffs.

## Refined Activity Landscape Features

The analysis of various landscape features can be refined further on the basis of conditional probabilities. Since a compound can have high probability for either category 1a or 1b and also for either category 2a or 2b, eight feature (and feature combination) categories can be defined for SAR-relevant activity landscape regions, as reported in **Table 1**. Utilizing the second conditioning, probabilities can be assigned to a compound to form similarity cliffs, smooth pairs, activity cliffs or different combinations of these features. For instance,

category 1a in **Table 1** represents compounds having a high probability to form similarity cliffs with similarly potent compounds.

**Table 1. Activity landscape feature probability classification**

| category | type | significance criterion | activity landscape feature probabilities |
|---|---|---|---|
| 0 | - | - | no significance |
| 1 | 1a | $P(\tilde{S}^{<}\|\tilde{A}^{<}, V_k) > P_T(\tilde{S}^{<}\|\tilde{A}^{<}, V_k)$ | similarity cliffs likely |
| 2 | 1b | $P(\tilde{S}^{\geq}\|\tilde{A}^{<}, V_k) > P_T(\tilde{S}^{\geq}\|\tilde{A}^{<}, V_k)$ | smooth pairs likely/similarity cliffs unlikely |
| 3 | 2a | $P(\tilde{A}^{<}\|\tilde{S}^{\geq}, V_k) > P_T(\tilde{A}^{<}\|\tilde{S}^{\geq}, V_k)$ | smooth pairs likely/activity cliffs unlikely |
| 4 | 2b | $P(\tilde{A}^{\geq}\|\tilde{S}^{\geq}, V_k) > P_T(\tilde{A}^{\geq}\|\tilde{S}^{\geq}, V_k)$ | activity cliffs likely |
| 5 | 1a, 2a | $P(\tilde{S}^{<}\|\tilde{A}^{<}, V_k) > P_T(\tilde{S}^{<}\|\tilde{A}^{<}, V_k)$ and $P(\tilde{A}^{<}\|\tilde{S}^{\geq}, V_k) > P_T(\tilde{A}^{<}\|\tilde{S}^{\geq}, V_k)$ | similarity cliffs likely/activity cliffs unlikely |
| 6 | 1a, 2b | $P(\tilde{S}^{<}\|\tilde{A}^{<}, V_k) > P_T(\tilde{S}^{<}\|\tilde{A}^{<}, V_k)$ and $P(\tilde{A}^{\geq}\|\tilde{S}^{\geq}, V_k) > P_T(\tilde{A}^{\geq}\|\tilde{S}^{\geq}, V_k)$ | similarity cliffs likely/activity cliffs likely |
| 7 | 1b, 2a | $P(\tilde{S}^{\geq}\|\tilde{A}^{<}, V_k) > P_T(\tilde{S}^{\geq}\|\tilde{A}^{<}, V_k)$ and $P(\tilde{A}^{<}\|\tilde{S}^{\geq}, V_k) > P_T(\tilde{A}^{<}\|\tilde{S}^{\geq}, V_k)$ | similarity cliffs unlikely/activity cliffs unlikely |
| 8 | 1b, 2b | $P(\tilde{S}^{\geq}\|\tilde{A}^{<}, V_k) > P_T(\tilde{S}^{\geq}\|\tilde{A}^{<}, V_k)$ and $P(\tilde{A}^{\geq}\|\tilde{S}^{\geq}, V_k) > P_T(\tilde{A}^{\geq}\|\tilde{S}^{\geq}, V_k)$ | similarity cliffs unlikely/activity cliffs likely |

Different activity landscape feature categories are reported. Categories consist of single features or combinations of features. For each category, the significance criterion is given and the corresponding landscape feature probabilities are described. *(taken from Vogt et. al[6])*

The significance criterion is satisfied if the corresponding probability exceeds the respective threshold. Likewise, compounds in category 2b have a high probability to form activity cliffs with immediate structural neighbors. Compounds that are likely to form smooth pairs can be further differentiated depending on

their inability to form similarity cliffs (category 1b) or activity cliffs (category 2a).

Four additional categories can be identified using combinations of these features, given that a compound has similar activity or is structurally related to other compounds. For example, combined category 1a-2a characterizes compounds that are likely to form similarity cliffs and unlikely to form activity cliffs. The opposite case is accounted for by category 1b-2b, i.e. compounds that are likely to form activity cliffs but unlikely to form similarity cliffs.

The per-compound conditional probabilities as well as their corresponding combinations calculated for all the compounds in a data set provide a refined view of the activity landscape features and their distribution. These are derived using full or partial memberships to different regions in the local SAS map [6].

# Applications

## Feature Probabilities and Thresholds

Conditional feature probabilities were calculated for 139 different activity annotated compound sets obtained from BindingDB [15]. The corresponding thresholds, i.e. $P_T(\tilde{S}^{<}|\tilde{A}^{<}, V_k)$, $P_T(\tilde{S}^{\geq}|\tilde{A}^{<}, V_k)$, $P_T(\tilde{A}^{<}|\tilde{S}^{\geq}, V_k)$, $P_T(\tilde{A}^{\geq}|\tilde{S}^{\geq}, V_k)$, were estimated from the combined conditional probabilities of all 139 data sets.

## Compound Assignment

For each compound, assignment to a category was performed according to the criteria outlined in **Table 1** if the corresponding conditional probability or a combination of conditional probabilities exceeded the respective threshold. Few categories reported in **Table 1** are mutually exclusive. For instance, a compound cannot be assigned to categories 2a and 2b ( i.e. *smooth pairs likely/activity cliffs unlikely* and *activity cliffs likely*). Other categories are not exclusive and give rise to combined categories.

For example, a compound can be assigned to categories 1a and 2b (i.e. *similarity cliffs likely* and *activity cliffs likely*), thus forming the combined category 1a-2b (i.e. *similarity cliffs likely/activity cliffs likely*). If a compound exceeded

threshold values for single as well as a combination of features, it was assigned to the combination.

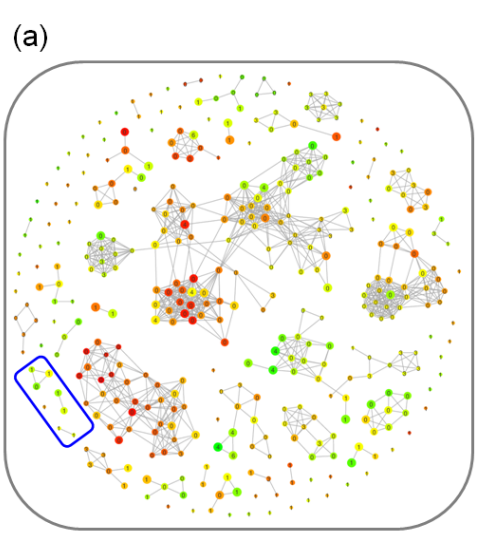
## Visualization and Exemplary Results

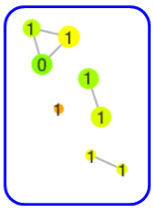
SAS maps can be utilized to visualize and characterize activity landscape features as they provide a basis for the analysis described herein. However, SAS maps are not suitable for analyzing feature probabilities of individual compounds as these represent compound pairs. Network-like similarity graph (NSG), a compound network-based activity landscape representation was utilized for the visualization of feature probabilities and compound assignment as the activity landscape is resolved at the level of individual compounds instead of compound pairs [16].

In an NSG representation, compounds are depicted as nodes colored according to their potency values from green (low potency) over yellow to red (high potency) and edges denote similarity relationships. Compound pairs are connected by an edge if their calculated ECFP4 Tc value is equal to 0.55 or greater. In addition, nodes are scaled in size according to compound discontinuity scores [16, 17].

Therefore, the larger the discontinuity introduced by the compound, the larger the node. Activity cliffs represent extreme forms of SAR discontinuity in an activity landscape [4]. Accordingly, in an NSG, the most prominent activity cliffs present in a data set are displayed as combinations of large red and green nodes connected by edges.

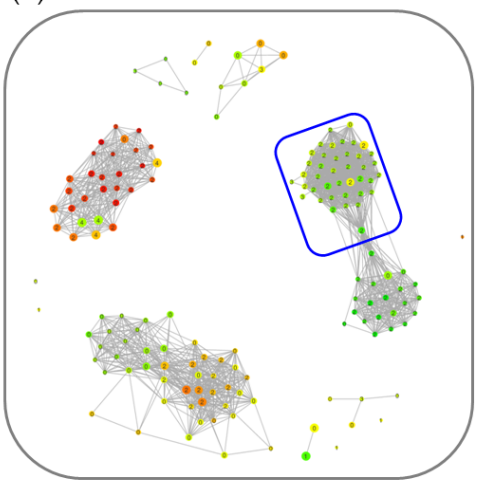
It should also be noted that the 2D arrangements of compounds and clusters in an NSG have no chemical meaning. Instead, the placement of compounds and the distances between them are determined by a graphical layout algorithm such that densely connected subsets of similar compounds are separated for clarity. In order to visualize compound feature probability information, nodes are annotated with the assigned categories according to **Table 1**.

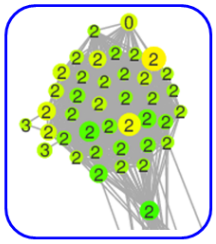


Serotonin receptor 7 (1521)

| Category | Count |
|---|---|
| 1 | 117 |
| 2 | 0 |
| 3 | 36 |
| 4 | 8 |
| 5 | 0 |
| 6 | 2 |
| 7 | 0 |
| 8 | 0 |



Endothelin receptor et-b (4513)

| Category | Count |
|---|---|
| 1 | 5 |
| 2 | 93 |
| 3 | 5 |
| 4 | 4 |
| 5 | 0 |
| 6 | 0 |
| 7 | 0 |
| 8 | 0 |

**Figure 1: Representative data sets.** Complete NSG representation for data sets containing (a) serotonin receptor 7 and (b) endothelin receptor et-b ligands are shown. Compound subsets in these representations are highlighted in blue. Enlarged views of these regions are provided and the numbers of compounds in various categories are reported. *(adapted from Vogt et. al[6])*

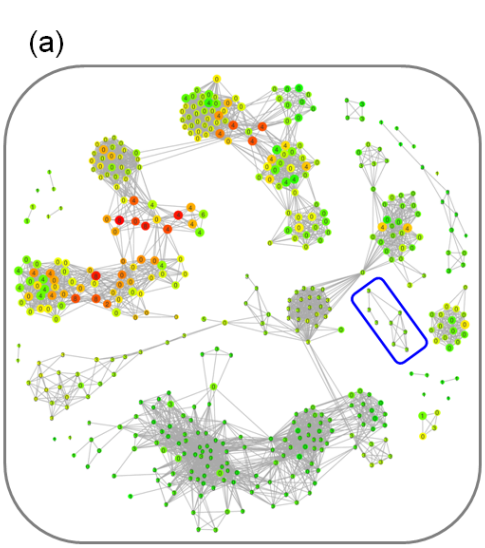
Following the systematic derivation of feature probabilities for individual compounds comprising all 139 data sets, significant differences were observed in conditional feature probability distribution. A number of data sets were identified that contained compounds belonging to one or two feature categories while others were dominated by compounds belonging to diverse categories without obvious preferences.

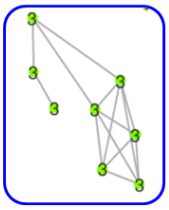
Varying numbers of compounds belonging only to categories *1* to *4* were found in 27 data sets but no data set contained compounds belonging only to feature combinations (categories *5-8*). Furthermore, it was also observed that data sets typically contained different numbers of compounds with probabilities lower than the respective thresholds (category *0*). These compounds were less likely to yield interpretable SAR information.

Selected data sets focusing on each of the eight feature categories (*1-8*) according to **Table 1** have been discussed further and the distribution of compounds within these categories has been reported. These exemplary compound sets demonstrate the variety of distributions observed and focus on individual feature categories. Complete NSGs for these data sets with enlarged highlighted sections for a detailed inspection has been utilized in the following as illustrated in **Figures 1-4**.

The serotonin receptor 7 data set in **Figure 1a** consists of 117 compounds that have a high probability to form similarity cliffs (category *1*). Several of these compounds have intermediate potencies and are structurally similar to a limited number of other compounds. Additionally, many category *0* compounds are also present in this data set. The conditional probabilities for these compounds do not reach their corresponding threshold values, including one within the selected subset. Thus, only limited SAR information content can be obtained from this structurally diverse ligand set.

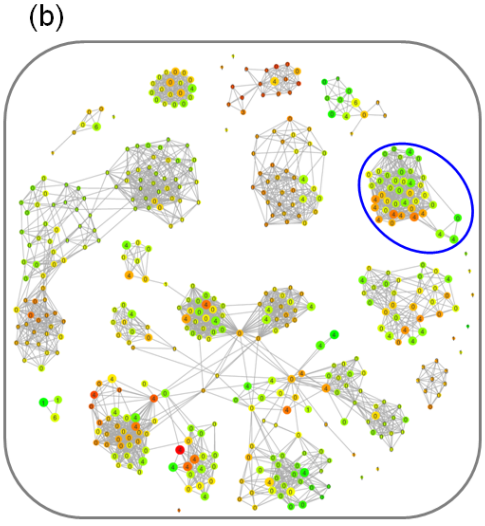
By contrast, majority of compounds in the endothelin receptor data set as seen in **Figure 1b** belong to category *2*. These compounds are likely to participate in the formation of smooth pairs and unlikely to form similarity cliffs. The NSG for this data set is dominated by densely connected compound clusters, consistent with this observation.

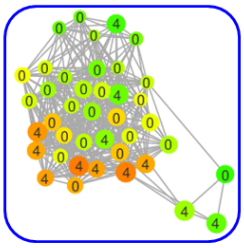


Melanocortin receptor 3 (5428)

| Category | Count |
|---|---|
| 1 | 26 |
| 2 | 1 |
| 3 | 150 |
| 4 | 31 |
| 5 | 1 |
| 6 | 1 |
| 7 | 0 |
| 8 | 5 |



Melatonin receptor 1b (1459)

| Category | Count |
|---|---|
| 1 | 20 |
| 2 | 0 |
| 3 | 8 |
| 4 | 55 |
| 5 | 0 |
| 6 | 3 |
| 7 | 0 |
| 8 | 0 |

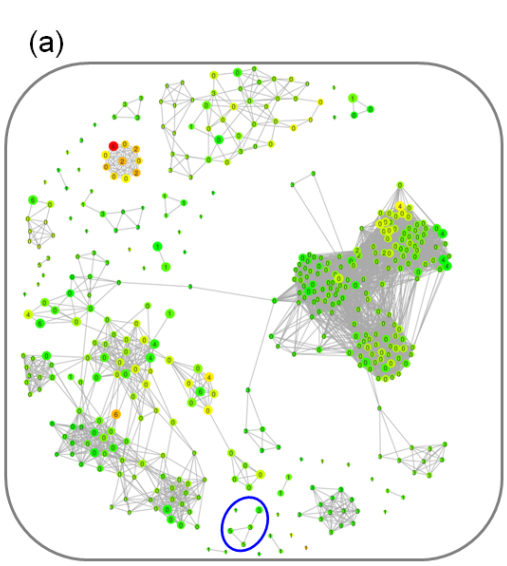
**Figure 2: Representative data sets.** Ligand sets for (a) melanocortin receptor 3 and (b) melatonin receptor 1b are shown according to the global versus local view used in Figure 1. *(adapted from Vogt et. al[6])*

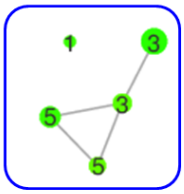
Analogously, several compounds that are likely to form smooth pairs but unlikely to form activity cliffs (category *3*) are observed in the data set containing melanocortin receptor 3 ligands as shown in **Figure 2a**. The graphical representation is characterized by structurally distinct subsets of compounds with different SAR information content.

The melatonin receptor 1b data set contains 55 compounds that are most likely to form activity cliffs (category *4*). In the NSG and the densely connected compound subset displayed in **Figure 2b**, prominent activity cliffs indicated by combinations of large red and green nodes can be clearly identified. The data set also contains several category *0* compounds of intermediate potencies. These compounds illustrated as nodes of small size introduce very little SAR discontinuity, although they are structurally related to activity cliffs. Category *4* and *0* compounds can not be differentiated on the basis of graphical representation alone. In such instances, compounds having a significant probability to form activity cliffs can only be distinguished from others by taking into account their conditional feature probabilities, thereby refining the activity landscape view.

The data set containing plasmin inhibitors in **Figure 3a** includes compounds belonging to category *5* that are likely to form similarity cliffs and unlikely to form activity cliffs. A small subset of category *5* compounds connected with those annotated as category *3*, i.e. compounds likely to form smooth pairs but unlikely to participate in activity cliff formation, is highlighted in the NSG. The graphical representation does not support the distinction between these compounds which can only be made on the basis of conditional probability calculations.

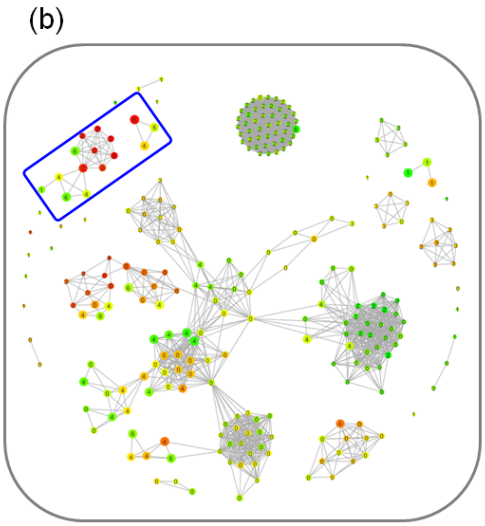
The cathepsin k inhibitor set reported in **Figure 3b** consists of eight compounds belonging to category *6* with significant probabilities for the formation of similarity cliffs and activity cliffs. The selected cluster includes four of these inhibitors, three of which have relatively low potency. Two compounds within the subset form an activity cliff but are structurally distinct from the other two as indicated by the absence of edges.

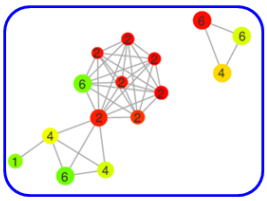



**Plasmin (1479)**

| Category | Count |
|---|---|
| 1 | 49 |
| 2 | 6 |
| 3 | 48 |
| 4 | 9 |
| 5 | 4 |
| 6 | 4 |
| 7 | 0 |
| 8 | 0 |




**Cathepsin k (1730)**

| Category | Count |
|---|---|
| 1 | 21 |
| 2 | 59 |
| 3 | 18 |
| 4 | 27 |
| 5 | 0 |
| 6 | 8 |
| 7 | 0 |
| 8 | 1 |

**Figure 3: Representative data sets.** Inhibitor sets for (a) plasmin and (b) cathepsin k data sets are shown. The arragement is consistent with the scheme used in Figure 1. *(adapted from Vogt et. al[6])*

In addition, the weakly potent cliff partner participates in the formation of similarity cliffs with the remaining three compounds.

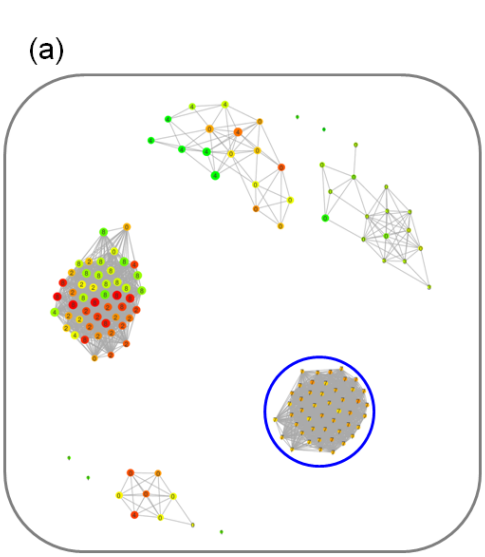
This information is evident from the graphical representation. The two activity cliff partners are also connected to a category *4* compound that is likely to form activity cliffs. Furthermore, two compounds within the selection also form activity cliffs with compounds having a high probability to form smooth pairs (category *2*) that exhibit chemical and activity similarity. Therefore, this compound subset is an example of a highly differentiated SAR micro environment that can be identified on the basis of conditional probabilities.

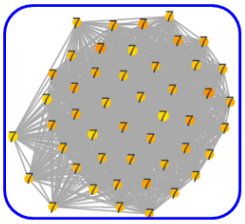
The data set composed of furin inhibitors in **Figure 4a** contain 47 compounds that are highly unlikely to participate in the formation of similarity cliffs and activity cliffs. The densely connected cluster highlighted in **Figure 4a** contains all of these compounds. In addition, these inhibitors have comparable potencies. Compounds within the subset belong to category *7*, i.e. they are structurally dissimilar to others in the data set and form smooth pairs with each other.

Finally, as seen in **Figure 4b**, the cholecystokinin-1 receptor data set predominantly consists of compounds belonging to category *8* that are unlikely to form similarity cliffs but likely to be involved in activity cliff formation.

Additionally, the subset also contains small numbers of category *0*, *2* and *4* compounds. Many category *8* compounds in the cluster produce multiple activity cliffs of large magnitude. Two weakly potent category *4* compounds also participate in cliff formation with category *8* compounds in this cluster. Overall, the cluster represents a rich source of SAR-relevant information as indicated by the presence of many large red and green nodes.

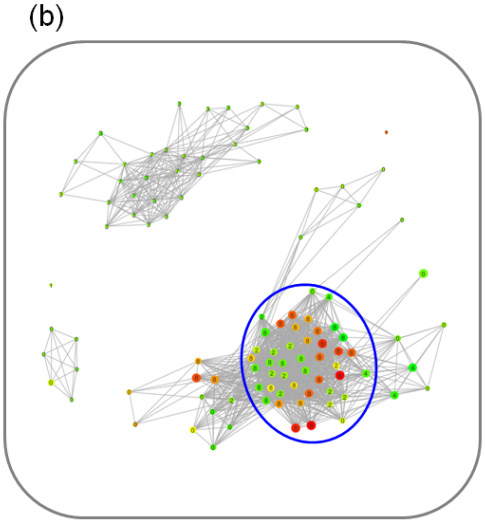
Hence, the examples shown in **Figures 1-4** clearly indicate that conditional activity landscape feature probabilities provide a refinement to the landscape views and aid in the differentiation of active compounds with respect to the associated SAR information.

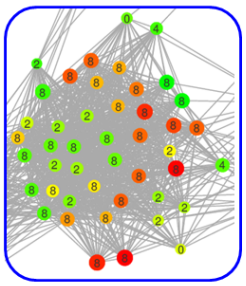


Furin
(3900)

| Category | Count |
|---|---|
| 1 | 5 |
| 2 | 20 |
| 3 | 5 |
| 4 | 12 |
| 5 | 0 |
| 6 | 0 |
| 7 | 47 |
| 8 | 23 |



Cholecystokinin-1
receptor (2552)

| Category | Count |
|---|---|
| 1 | 2 |
| 2 | 12 |
| 3 | 24 |
| 4 | 4 |
| 5 | 0 |
| 6 | 0 |
| 7 | 5 |
| 8 | 33 |

**Figure 4: Representative data sets.** Data sets containing (a) furin and (b) cholecystokinin-1 receptor ligands are shown according to the scheme used in Figure 1. *(adapted from Vogt et. al[6])*

# Summary

A novel approach for the assignment of activity landscape features at the level of individual compounds has been introduced and the derivation of conditional feature probabilities using fuzzy boundaries to delineate the different activity landscape regions formed the basis of this methodology. Utilization of fuzzy boundaries results in partial feature memberships for compounds and balances possible boundary effects. Conditional probabilities have been obtained from pairwise chemical similarity and activity difference relationships between compounds and the frequency derived feature analysis has been carried out for individual compounds in a data set. Local SAS maps have been introduced for this purpose. The resulting per-compound conditional feature probabilities provides a conceptual advance in the analysis of activity landscape. Furthermore, conditional probability calculations have made it possible to derive eight different feature categories from the existing three compound pair-based SAR-relevant landscape regions, i.e. activity cliffs, smooth pairs and similarity cliffs. Assignment of compounds to these categories allows their further distinction in local SAR environments. These localized SAR environments were difficult to interpret on the basis of graphical landscape representations alone, despite taking into account numerical SAR discontinuity measures. Herein, the emphasis has been to demonstrate the differentiation of SAR micro environments using compound conditional probabilities in these graphical representations. Thus, the conditional probability calculations and the ensuing categorization scheme further refine the current activity landscape views and aid in the systematic SAR analysis at level of individual compounds.

The study reported herein has been published in reference [6] of this chapter. My contributions to this study have been the systematic calculation of conditional feature probabilities for the 139 data sets and the generation of exemplary NSGs.

# References


[1] Wassermann A. M., Wawer M., Bajorath J. Activity landscape representations for structure-activity relationship analysis. *J. Med. Chem.*, **2010**, 53, 8209-8223.

[2] Stumpfe D., Bajorath J. Methods for SAR visualization. *RSC Adv.*, **2012**, 2, 369-378.

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

[4] Stumpfe D., Bajorath, J. Exploring activity cliffs in medicinal chemistry. *J. Med. Chem.*, **2012**, 55, 2932-2942.

[5] Iyer P., Stumpfe D., Vogt M., Bajorath J., Maggiora G. M. Activity landscapes, information theory, and structure-activity relationships. *Mol. Inf.*, **2013**, 32, 421-430.

[6] Vogt M., Iyer P., Maggiora G. M. and Bajorath J. Conditional probabilities of activity landscape features for individual compounds. *J. Chem. Inf. Model.*, **2013**, 53, 1602-1612.

[7] Shanmugasundaram V., Maggiora G. M. Characterizing property and activity landscapes using an information-theoretic approach. *222$^{nd}$ ACS National Meeting.*, **2001**, Division of Chemical Information, Abstract no. 77.

[8] Willett P. Searching techniques for databases of two- and three-dimensional structures. *J. Med. Chem.*, **2005**, 48, 4183-4199.

[9] Perez-Villanueva J., Santos R., Hernandez-Campos A., Giulianotti M. A., Castillo R., Medina-Franco J. L. Structure-activity relationships of benzimidazole derivatives as anti-parasitic agents: dual-activity difference (DAD) maps. *Med. Chem. Commun.*, **2011**, 2, 44-49.

[10] 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.

[11] Rogers D., Hahn M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, **2010**, 50, 742-754.

[12] 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.

[13] Wawer M., Bajorath J. Similarity-Potency Trees: a method to search for SAR information in compound data sets and derive SAR rules. *J. Chem. Inf. Model.*, **2010**, 50, 1395-1409.

[14] Zimmermann, H. -J. Fuzzy Set Theory. *WIRES Computational Statistics*, **2010**, 2, 317-332.

[15] Liu T., Lin Y., Wen X., Jorissen R. N., Gilson M. K. Binding-DB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.*, **2007**, 35, D198-D201.

[16] 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.

[17] Peltason L., Bajorath J. SAR Index: quantifying the nature of structure-activity relationships. *J. Med. Chem.*, **2007**, 50, 5571-5578.

Systematic SAR analyses using activity landscape representations focus on the distribution of molecular similarities and activity data associated with bioactive compounds. In the next chapter, the existing framework of landscape modeling has been modified to integrate mechanism of action information of receptor ligands.

# Chapter 4

# 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

## Introduction

An important objective of SAR analyses is to examine the associations that exist between chemical structures of bioactive compounds and their activity. Biological targets may include different enzymes or receptors. Receptors are important targets for therapeutic drugs. Pharmacological theory postulates that the types of functional responses produced by chemical agents are determined by their interactions with target receptors. Thus, ligands can be classified

as full agonists, partial agonists, antagonists or inverse agonists, depending on their mechanism of action. However, the mechanisms of action of ligands have not been considered during typical SAR analyses thus far. A modified activity landscape model has been introduced that accounts for the mechanistic information of receptor ligands. This resulted in a graphical network representation that combined systematic similarity and potency relationships in addition to mechanism-related information. Simultaneous analysis of SAR and mechanism of action is helpful in the identification of structurally similar compounds with different mechanistic behavior. Following the inspection of such ligands, structural changes that lead to "mechanism hops" can be inferred.

JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

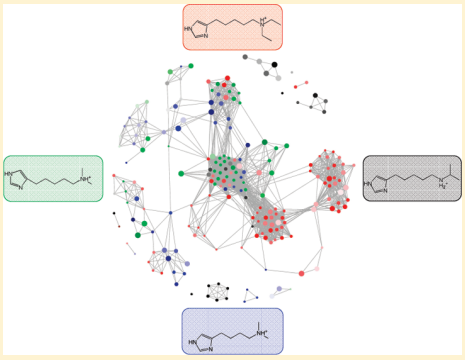
pubs.acs.org/jcim

# 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

Preeti Iyer, 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:** Receptor ligands might act as agonists, partial agonists, inverse agonists, or antagonists and it is often difficult to understand structural modifications that alter the mechanism of action. In order to compare ligands that are active against a given receptor but have different mechanisms of action, we have designed molecular networks that mirror similarity relationships and incorporate both mechanism of action information and mechanism-specific SAR features. These network representations make it possible to systematically evaluate relationships between different types of receptor ligands and identify communities of structurally very similar ligands with different mechanisms. From a series of such ligands, structural modifications can often be deduced that lead to "mechanism hops".



## INTRODUCTION

Pharmacological effects of receptor ligands arise from diverse mechanisms of action.[1,2] For example, ligands might act as agonists, partial agonists, inverse agonists, or antagonists. In general terms, the mechanisms of these types of ligands can be defined as follows: (1) An *agonist* binds to the physiological ligand binding site of a receptor and activates it. (2) An *antagonist* blocks this binding site and thereby prevents receptor activation and signaling. (3) A *partial agonist* also competes with the natural ligand but does not fully activate the receptor. (4) An *inverse agonist* stabilizes an inactive conformation of a receptor and thereby prevents activation and signal transduction.

Mechanistic effects are in general closely linked to different conformational states, or conformational ensembles, of receptors and their ability to interact with effector proteins. G protein-coupled receptors (GPCRs), for which receptor−ligand interactions are just beginning to be understood at the molecular level of detail,[3,4] are prime examples of receptors that engage in highly complex mechanisms controlling functional effects.[3] Differences in the cellular context of GPCR-ligand interactions are also known to alter pharmacological profiles of ligands,[5] giving rise to an intricate network of factors that ultimately determine pharmacological effects. From a medicinal chemistry perspective, it is often difficult to differentiate between different modes of action of GPCR ligands and identify structural determinants of a specific mechanism.[6] Consequently, approaches that help to reveal structural features that influence or determine the mechanism of action of GPCR or other receptor ligands are particularly attractive for medicinal chemistry applications.

In recent years, molecular network representations have been increasingly utilized to systematically account for ligand-target interactions and predict targets of active compounds[7] or analyze structure−activity relationships (SARs).[8] Such molecular networks make it possible to analyze large data sets using a consistent representation frame and often provide graphical access to unexpected ligand-target interactions[7] or complex SAR features.[8] As such, they complement more traditional approaches to analyze ligand-target interactions or SARs.

Given the often complex mechanistic spectrum of receptor ligands, as discussed above, we have been interested in the design of molecular network representations that help to compare ligands with different mechanisms. Therefore, we have generated similarity-based compound networks that incorporate mechanistic and SAR information. These graphical representations make it possible to identify compounds that are related to each other but act by different mechanisms and determine structural features that lead to "mechanism hopping".

## MATERIALS AND METHODS

**Compound Data Sets.** For five different GPCRs, ligand sets were collected from the ChEMBL database.[9] These data sets were assembled to contain ligands having different mechanisms of action including agonists, partial agonists, inverse agonists, and antagonists. The composition of these compound sets is

**Received:** March 21, 2011
**Published:** May 07, 2011

summarized in Table 1. The sets contained between 148 (AM1) and 307 (AA1) ligands. In two of five cases, S1A and AM1, no inverse agonists were available. All mechanistic annotations for ligands taken from ChEMBL were extracted from original literature sources. Molecules with both agonist and partial agonist annotations were classified as *partial agonists*, and molecules designated as full agonists were classified as *agonists*. In addition, ligands with only inverse agonist or both antagonist and inverse agonist annotations were classified as *inverse agonists*, owing to the observation that compounds that are apparent GPCR antagonists act in many cases by an inverse agonist mechanism.[6] Ligands designated only as antagonists were classified as *antagonists*. As potency annotations, only $K_i$ (or p$K_i$) were considered. If multiple $K_i$ values were reported for a ligand in original literature sources, the geometric mean was calculated to yield a final potency value.

**Network-like Similarity Graphs.** The network-like similarity graph (NSG) data structure[10] is a similarity-based molecular network representation that is annotated with additional information layers. Nodes are molecules and edges between them indicate pairwise similarity relationships. Nodes are color-coded using a continuous spectrum to reflect the potency distribution in a compound data set and scaled in size according to the distribution of per-compound discontinuity scores. The per-compound discontinuity score indicates the amount of local SAR discontinuity a compound introduces.[10] Hence, a compound that greatly differs in potency from its immediate structural neighbors makes a large contribution to local SAR discontinuity, and, accordingly, the corresponding node is large. NSGs are generated utilizing the Java implementation in the publicly available SARANEA program[11] and applying a graphical layout algorithm[12] that places densely connected compound subsets in close proximity and separates weakly interconnected regions (compound clusters) from each other. NSGs are usually generated for a set of compounds active (with different potencies) against a given target, i.e. a compound activity class.

**Mechanism-Based NSGs.** In order to compare receptors ligands with different mechanisms of action using a consistent representation frame, we designed an NSG variant that incorporates mechanism of action information. Therefore, the NSG data structure was modified in different ways, as further detailed in the Results and Discussion section. Pairwise similarity relationships were calculated using the stereochemistry-sensitive Extended Connectivity Fingerprint with bond diameter 4 (ECFP4)[13] as implemented in Pipeline Pilot.[14] An ECFP4 Tanimoto coefficient (Tc) value of 0.4 was applied as the similarity threshold for edges between nodes. This ECFP4 Tc value roughly corresponds to a MACCS structural keys[15] Tc value of 0.8 and indicates the presence of compounds with visible structural similarity. Different from NSGs, nodes were not calculated by potency, but rather by mechanism, using the following color scheme: *agonists*, blue; *partial agonists*, green; *inverse agonists*, gray; *antagonists*, red. For each mechanistic class (compound subset), potency information was conveyed by shading, i.e. for each mechanism color, a continuous shade spectrum from transparent (lowest potency) to opaque (highest potency) was applied. These graphs were implemented in Java, further extending the SARANEA implementation,[11] and termed *Mechanism-based NSGs* (*M-NSGs*). Table 2 summarizes the design elements of M-NSGs.

**Table 1. Receptor Ligand Sets**[a]

| | | | p$K_i$ | |
|---|---|---|---|---|
| target receptor | ligand mechanism | no. of compounds | maximum | minimum |
| adenosine A1 | agonist | 107 | 9.8 | 4.9 |
| receptor (AA1) | partial agonist | 54 | 8.7 | 5.3 |
| | antagonist | 94 | 9.5 | 6.0 |
| | inverse agonist | 52 | 9.4 | 4.2 |
| muscarinic | agonist | 26 | 8.6 | 3.6 |
| acetylcholine | partial agonist | 49 | 9.5 | 4.5 |
| receptor M1 (AM1) | antagonist | 73 | 10.0 | 4.5 |
| dopamine D2 | agonist | 40 | 9.0 | 5.6 |
| receptor (DD2) | partial agonist | 44 | 9.7 | 6.2 |
| | antagonist | 76 | 9.8 | 6.6 |
| | inverse agonist | 13 | 11.5 | 6.1 |
| histamine H3 | agonist | 44 | 9.6 | 4.9 |
| receptor (H3R) | partial agonist | 46 | 9.1 | 5.0 |
| | antagonist | 92 | 10.0 | 6.8 |
| | inverse agonist | 31 | 10.1 | 5.8 |
| serotonin 1a | agonist | 46 | 10.2 | 5.4 |
| receptor (S1A) | partial agonist | 78 | 10.1 | 4.9 |
| | antagonist | 63 | 9.4 | 6.4 |

[a] Receptor abbreviations are used in the text to designate ligand sets.

## RESULTS AND DISCUSSION

**Graph Design.** M-NSG generation involved different types of calculations, either for an entire ligand set or for each separate mechanism-based subset. First, the graph layout was computed for a complete ligand set after calculating pairwise Tanimoto similarity for all ligands, regardless of their mechanisms of action, thus providing the similarity-based compound network, in analogy to original NSGs. Then, however, similarity- and potency-based discontinuity scores were separately calculated for each ligand subset. On the basis of subset-specific discontinuity scores, the nodes of ligands sharing the same mechanism were scaled in size, hence providing SAR information for each mechanism-based subset. Node scaling is interpreted in the following manner: the larger a node, the higher the degree of SAR discontinuity the compound introduces; combinations of connected medium to large nodes represent discontinuous local SARs and combinations of small nodes continuous local SARs. Finally, mechanism and compound potency information was

**Table 2. M-NSG Design Elements**

| mechanism | color | node size | shading | edge | layout |
|---|---|---|---|---|---|
| agonist | blue | mechanism-specific per compound discontinuity score | potency range | Tanimoto similarity >0.4 | connectivity-based |
| partial agonist | green | | | | |
| antagonist | red | | | | |
| inverse agonist | gray | | | | |

incorporated through color and shade coding. Each ligand mechanism was assigned a specific color, and compounds sharing the same mechanism were shaded according to their potency level (within the potency range in the subset). Hence, combinations of the largest transparent and opaque nodes sharing the same color mark the most prominent activity cliffs that occur in each subset. Activity cliffs are structurally similar compounds with very different potency (representing the extreme form of SAR discontinuity).[8] On the basis of these design components, M-NSGs provide similarity and mechanism information across an entire receptor ligand set and, in addition, relative potency and SAR information for each mechanism-specific subset. Thus, M-NSGs also contain all the information provided by NSGs of individual compound sets sharing a specific mechanism of action. Because SAR information can also be obtained from NSGs of individual mechanism-based ligand subsets, we predominantly focus in the following on the exploration of mechanism hopping and underlying structural changes, for which M-NSGs are specifically designed.

**Ligand M-NSGs.** Figure 1 shows the M-NSG representations for the five receptor ligand sets in Table 1. The M-NSG of the AA1 ligand set in Figure 1a reveals a central graph component (region 1) and several other densely connected compound clusters. In many instances, these clusters mostly, or exclusively, consist of ligands having the same mechanism, e.g. antagonist clusters (red nodes). The central graph component contains partial (green), full agonists (blue), a few inverse agonists (gray), and many differently sized nodes. This indicates the presence of substantial SAR information for agonists in this region. Similarly, the mostly gray-scaled compound community (region 2) at the bottom in Figure 1a contains many inverse agonists with differently sized nodes and a few antagonists having small nodes. Although the AA1 M-NSG displays a notable clustering of compounds by mechanism, there are exceptions, in particular, a densely connected community of ligands with all four mechanisms of action (region 3). Such densely connected mechanistically heterogeneous ligand communities represent prime candidates for further analysis to explore the structural basis of mechanistic changes among similar ligands. In addition, mechanistically more homogeneous regions such as the central graph component in Figure 1a are a source of SAR information for compounds sharing the same or similar mechanisms.

The DD2 M-NSG in Figure 1b represents the smallest of the five data sets and is characterized by the presence of structurally diverse compounds. Structural diversity is mirrored by the low edge density in the graph. The most notable feature of the DD2 M-NSG is its largest ligand community in the center of Figure 1b (region 1) that also contains compounds with all four mechanisms of action.

By contrast, the H3R M-NSG in Figure 1c is characterized by the presence of a densely connected central graph component that essentially consists of four separate ligand communities that are connected via compound bridges. These include two antagonist communities (regions 1 and 2) with different node size and potency distributions and, in addition, two other communities that are characterized by distinct mechanistic heterogeneity (regions 3 and 4). In particular, the community in the center of Figure 1c (region 4) consists of very similar ligands that cover the entire spectrum of mechanisms and thus provides a focal point for further analysis.

Different from the ligand sets discussed so far the remaining AM1 and S1A sets in Figure 1d and 1e, respectively, do not contain inverse agonists. The AM1 M-NSG shows a notable clustering of different series of compounds by mechanism. However, in some cases, individual ligands with different mechanisms occur in an otherwise mechanistically homogeneous cluster including, for example, a weakly potent antagonist found in a partial agonist community (region 1 in Figure 1d), another single antagonist in an agonist/partial agonist community (region 2), and a weakly potent partial agonist within an antagonist community (region 3). Such observations might raise the question as to whether the mechanisms of these individual ligands located in an otherwise mechanistically homogeneous environment have been correctly identified and might thus suggest further experimental evaluation. In addition, another small community (region 4) shows a sequence of agonists with increasing potency where structural neighbors of potent agonists include partial agonists and a weakly potent antagonist, which represents another interesting and perhaps puzzling mechanistic pattern. Furthermore, another ligand community is encircled in Figure 1d (region 5) that contains multiple agonists, partial agonists, and antagonists with different potencies.

The S1A M-NSG in Figure 1e also contains both mechanistically homogeneous and heterogeneous ligand communities. For example, two densely connected ligand communities are observed (regions 1 and 2) that each comprise multiple partial agonists and multiple antagonists. Furthermore, region 3 in the S1A M-NSG contains a pair of compounds representing an agonist/antagonist hop and a small community of structurally related ligands including an antagonist, an agonist, and two partial agonists.

**Mechanism Hopping.** Selected mechanistically heterogeneous ligand communities were analyzed in detail to explore structural modifications that lead to mechanistic changes. M-NSG regions shown in Figure 2 are labeled with red numbers in Figure 1. In each community, a series of analogs were identified that revealed structural changes altering their mechanisms of action.

The series of AA1 ligands in Figure 2a includes agonists, partial agonists, inverse agonists, and an antagonist. All of these ligands are analogs and only distinguished by different substituents at the same site, i.e. a meta-position of the phenyl ring. An agonist (ligand 1) contains a hydroxyl group at this position. Ligands with a methoxy or methyl group (2 and 3) are partial agonists, and the same mechanism is observed for a fluorine substituent (4). However, changing the fluorine atom to a difluoromethylether group converts an agonist (4) into an antagonist (5). Moreover, changing this group to either a trifluoromethylether or trifluoromethyl substituent generates inverse agonists (6 and 7). Thus, ligands taken from a mechanistically heterogeneous region in the AA1 M-NSG reveal substitutions at a single site that alter the mechanism of action in different ways.

The largest ligand community in the DD2 M-NSG contains three tetraline analogs with different mechanisms that are shown in Figure 2b. These analogs include an agonist (ligand 1), antagonist (2), and inverse agonist (3). Here substitutions at multiple positions in both rings and different stereochemistry at the aliphatic ring distinguish the inverse agonist and antagonist from the agonist.

Five ligands from a mechanistically highly heterogeneous region of the H3R M-NSG shown in Figure 2c include an agonist, a partial agonist, two antagonists, and an inverse agonist. Their structures are distinguished by the length of the aliphatic linker between the imidazole ring and the terminal amine and, in

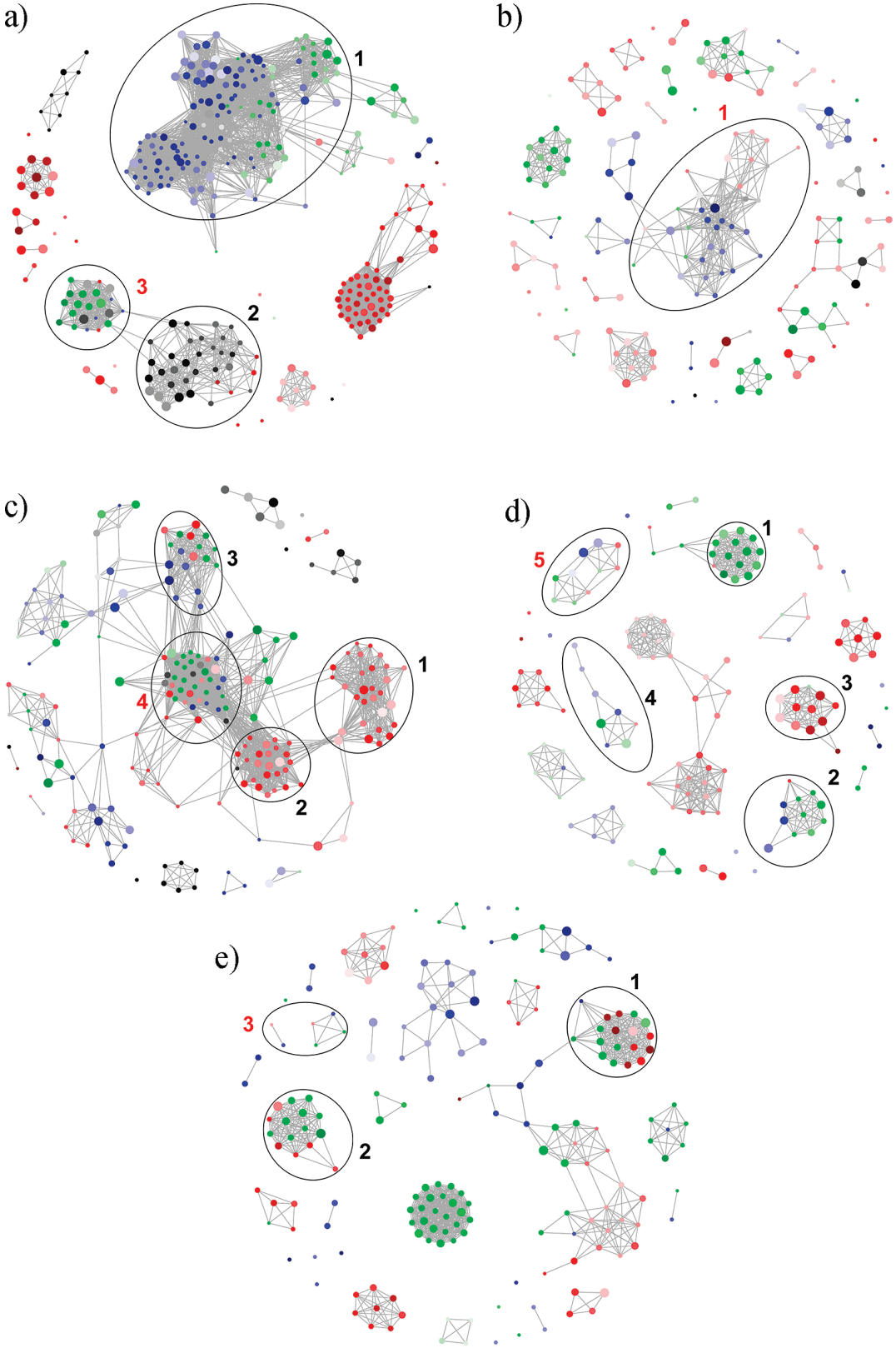



**Figure 1.** M-NSGs. Graph representations are shown for each complete ligand set according to Table 1 (agonists, blue; partial agonists, green; inverse agonists, gray; antagonists, red). Regions of the graph discussed in the text are encircled and numbered. Regions with red numbers are shown in detail in Figure 2. (a) AA1, (b) DD2, (c) H3R, (d) AM1, (e) S1A. In each M-NSG, selected ligand communities are indicated.


dx.doi.org/10.1021/ci2001378 |*J. Chem. Inf. Model.* 2011, 51, 1281–1286

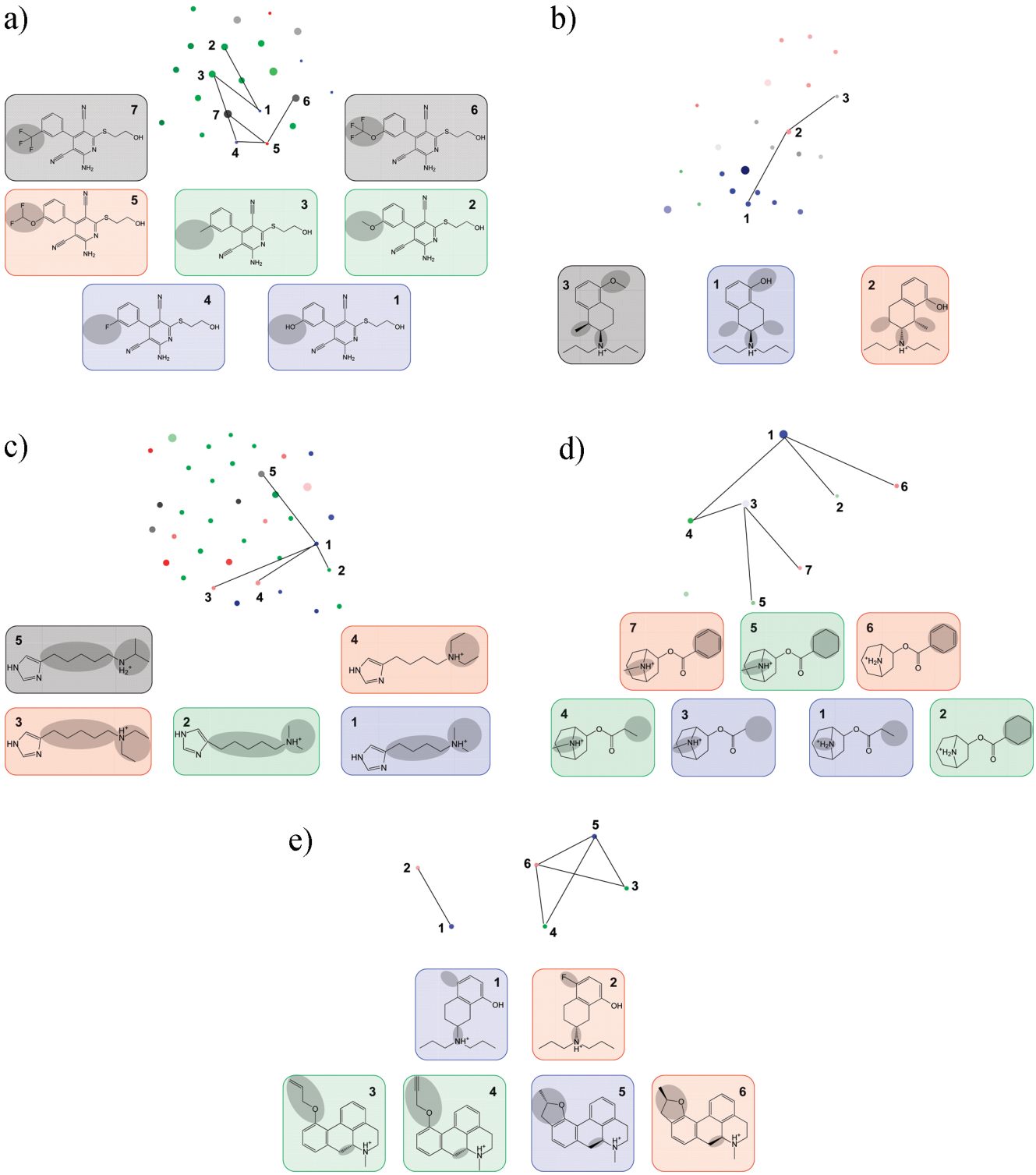



**Figure 2.** Ligand communities. For each data set, selected ligand communities are displayed (communities with red numbers in Figure 1). For each community, a series of ligands is shown that include mechanism hops. In the community graphs, nodes representing pairs of these ligands that constitute mechanism hops are connected by black edges. Compounds and the corresponding nodes are numbered. Structural modifications in ligands are highlighted, and they are displayed on a background representing their node colors (mechanisms). (a) AA1, (b) DD2, (c) H3R, (d) AM1, (e) S1A.

addition, by substitutions at this amino group. Comparison of ligands 1−4 reveals that the linker length is not responsible for agonistic versus antagonistic effects. Rather, the introduction of an N,N-diethylamino group (in ligands 3 and 4) generates antagonists. If the N,N-diethylamine is changed to an N-tertiary butyl group, an inverse agonist is obtained. Clearly, mechanistic changes can in this case be attributed to the substitutions at the terminal amino group.


dx.doi.org/10.1021/ci2001378 |*J. Chem. Inf. Model.* 2011, 51, 1281–1286

The series of AM1 ligands in Figure 2d contains five agonists or partial agonists (ligands 1−5) and two antagonists (6 and 7). Both agonists and antagonists contain two forms of a bridged hetereoaliphatic ring, which can thus not be responsible for changes in the mechanism of action. By contrast, mechanism hopping from agonists to antagonists is caused by the introduction of a phenyl substituent at the ester moiety, instead of a small aliphatic group or a cyclohexyl ring. Comparisons of ligands 2 and 6 and of ligands 5 and 7 reveal that the replacement of the cyclohexyl by the benzene ring, i.e. the introduction of an aromatic ring at this position, converts partial agonists into antagonists, another well-defined chemical modification.

In Figure 2e, a pair of S1A ligands is shown (1 and 2) that are also tetraline derivatives, similar to the ligands in Figure 2b. This chemotype is active against both dopamine and serotonin receptors. The two ligands in Figure 2e represent a mechanism hop from an agonist to an antagonist. In both analog series in Figure 2b and 2e, we observe that the stereochemistry at the nitrogen is a differentiating feature between agonists and antagonists. Ligands 1 and 2 in Figure 2e are further distinguished by a fluorine substituent. In addition, Figure 2e also shows another series of S1A ligands (3 to 6) containing a condensated ring system as their core structure. Here different core ring stereochemistry is observed as well as modifications at a phenyl moiety. The two partial agonists (ligands 3 and 4) differ from the full agonist (5) in a ring stereoisomer and a methyl-tetrahydrofuran fused to the phenyl moiety that is only present in the full agonist. However, the agonist and the antagonist (6) in this series are nearly identical; they only differ in the stereoisomer of the methyl substituent at the tetrahydrofuran ring. Thus, in this case, a subtle stereochemical difference involving a single methyl group in chemically complex and rigid receptor ligands triggers a change in the mechanism of action from an agonist to an antagonist.

## CONCLUSIONS

Herein we have introduced a graphical analysis tool that incorporates molecular mechanism of action information. The M-NSG data structure makes it possible to graphically analyze sets of receptor ligands with different mechanisms of action and identify mechanistically heterogeneous communities of structurally similar compounds. In the M-NSG implementation, nodes are directly associated with compound structures for interactive display. Hence, compound subsets can be easily selected and further analyzed to explore structural modifications that might lead to mechanistic changes. M-NSG analysis has been carried out for five sets of GPCR ligands acting by three or four different mechanisms. In a number of instances, well-defined structural changes were identified in analog series prioritized on the basis of M-NSG analysis that distinguished between ligands with different mechanisms. For closely related receptors (e.g., isoforms), it might also be possible to pool active compounds by mechanism and study these sets in M-NSGs in order to search for structural changes that might be responsible for similar mechanistic effects across multiple receptors.

## AUTHOR INFORMATION

### Corresponding Author

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

## REFERENCES

(1) Kenakin, T. Principles: Receptor Theory in Pharmacology. *Trends Pharmacol. Sci.* **2004**, *25*, 186–192.

(2) Zhu, B. T. Mechanistic Explanation for the Unique Pharmacologic Properties of Receptor Partial Agonists. *Biomed. Pharmacol.* **2005**, *59*, 76–89.

(3) Rosenbaum, D. M.; Rasmussen, S. G.; Kobilka, B. K. The Structure and Function of G Protein-Coupled Receptors. *Nature* **2009**, *459*, 356–363.

(4) Sprang, S. R. Binding the Receptor at Both Ends. *Nature* **2011**, *469*, 172–173.

(5) Nelson, C. P.; Challiss, R. A. 'Phenotypic' Pharmacology: the Influence of Cellular Environment on G Protein-Coupled Receptor Antagonist and Inverse Agonist Pharmacology. *Biochem. Pharmacol.* **2007**, *73*, 737–751.

(6) Greasley, P. J.; Clapham, J. C. Inverse Agonism or Neutral Antagonism at G Protein-Coupled Receptors: a Medicinal Chemistry Challenge worth Pursuing? *Eur. J. Pharmacol.* **2006**, *553*, 1–9.

(7) Keiser, M. J.; Irwin, J. J.; Shoichet, B. K. The Chemical Basis of Pharmacology. *Biochemistry* **2010**, *49*, 10267–10276.

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

(9) *ChEMBL*; European Bioinformatics Institute (EBI): Cambridge, 2010. http://www.ebi.ac.uk/chembl/ (accessed March 2, 2011).

(10) Wawer, M.; Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Structure-Activity Relationship Anatomy by Network-like Similarity Graphs and Local Structure-Activity Relationship Indices. *J. Med. Chem.* **2008**, *51*, 6075–6084.

(11) 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.

(12) Fruchterman, T. M. J.; Reingold, E. M. Graph Drawing by Force-directed Placement. *Software − Pract. Exper.* **1991**, *21*, 1129–1164.

(13) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(14) *Scitegic Pipeline Pilot*, Student ed.; Version 6.1; Accelrys, Inc.: San Diego, CA, 2007.

(15) *MACCS Structural Keys*; Symyx Software: San Ramon, CA, 2005.

## Summary

This chapter outlines a computational methodology to systematically analyze the mechanism of action information associated with ligand sets for target receptors in addition to their molecular similarity and activity distribution. Chemical similarity is accessed using molecular fingerprint representation during the generation of these network-based activity landscapes. This results in a graphical organization of both SAR and mechanism-specific content underlying various receptor ligand sets. Such combined views highlight ligand subsets with mechanistic homogeneity and heterogeneity. Further exploration of these heterogeneous compound clusters reveal chemical substitutions that introduce mechanistic transitions. In addition, SAR trends present in compounds with a given mechanism of action help to identify structural modifications that lead to improvements in potency.

A second activity landscape representation that incorporates mechanism of action information is introduced in the next chapter.

# Chapter 5

# Mechanism-based bipartite matching molecular series graphs to identify structural modifications of receptor ligands that lead to mechanism hopping

## Introduction

Identification of chemical substitutions that improve or adversely affect compound potencies help in deducing SAR rules necessary for compound optimization. Computational techniques based on molecular fingerprint representations require close examination of the 2D structures to identify such modifications. However, the matched molecular pair (MMP) paradigm helps to readily identify those structural changes that are favorably associated with activity. MMPs are formed by pairs of compounds that differ only by a single substructure exchange [1]. MMP-based similarity criterion has recently been applied in the design of a novel graphical tool with the objective of illustrating substructure relationships within compound data sets. This approach has been extended to account for mechanism of action information to highlight those chemical replacements that lead to mechanistic switches or mechanism hops with relative ease. Fur-

ther, organization of substructures in a hierarchy aids in the identification of increasingly smaller substructure changes involved in mechanism hopping.

[1] Kenny P. W., Sadowski J. Structure modification in chemical databases. In Oprea T. I. (Ed.), *Chemoinformatics in Drug Discovery*, Wiley-VCH, Weinheim, Germany, **2004**, 271-285.

MedChemComm

Cite this: *Med. Chem. Commun.*, 2012, **3**, 441

www.rsc.org/medchemcomm

CONCISE ARTICLE

# Mechanism-based bipartite matching molecular series graphs to identify structural modifications of receptor ligands that lead to mechanism hopping

**Preeti Iyer and Jürgen Bajorath***

***Received 9th November 2011, Accepted 22nd December 2011***
**DOI: 10.1039/c2md00281g**



The rationalization of structural features that distinguish between different mechanisms of action of ligands active against a given receptor is of high importance in medicinal chemistry and drug design. We have adapted a bipartite molecular network structure that organizes compound datasets on the basis of substructure relationships to incorporate mechanism of action information. The resulting data structure readily identifies subsets of ligands with different mechanisms of action that display well-defined structural relationships. From the structural subset organization of the graph, structural changes that lead to mechanism hopping (*i.e.*, a transition from one mechanism of action to another) can be directly selected, as demonstrated for different classes of receptor ligands. For medicinal chemistry applications, the ability to immediately access structural modifications that distinguish ligands having different mechanisms of action is a key aspect of the methodology introduced herein. The knowledge of substituents in receptor ligands that trigger mechanistic changes can be utilized for compound design.

## Introduction

Compounds that are active against a given receptor often display different mechanisms of action that ultimately lead to receptor activation or inactivation.[1–3] The mechanistic spectrum of receptor ligands and the ensuing functional effects are often more complex, and even more difficult to understand, than mechanisms of enzyme inhibitors, which might sterically block access to a catalytic site or act as transition state analogs,[4,5] or mechanisms of allosteric enzyme activators.[6] In medicinal chemistry, understanding structural features that determine the mechanisms of action of different types of receptor ligands represents a challenging task of central relevance.[2,3] For receptor ligands, it is often very difficult to discern the molecular basis of different mechanistic and functional effects.[2]

Mechanism of action studies are typically carried out in the context of compound structure–activity relationship (SAR) analysis. In order to extract SAR information from large and chemically diverse compound datasets, numerical SAR analysis functions and molecular network representations have increasingly been used.[7–9] Similarly, network representations have also been utilized to systematically account for ligand–target associations and explore the molecular basis of compound pharmacology.[10,11]

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

In order to further refine SAR investigations of receptor ligands, similarity-based compound networks might also be annotated with mechanistic information. For this purpose, network-like similarity graphs (NSGs)[12] have recently been employed. The NSG structure captures compounds as nodes that are connected by edges if pair-wise compound similarity reaches a pre-defined threshold level. Nodes and edges can be annotated with additional information. Typically, nodes are color-coded by compound potency to provide a basis for graphical SAR analysis.[12] However, in order to incorporate mechanistic information into these network representations, a molecular mechanism-based color code for nodes has also been introduced in the design of an NSG variant.[13] In this case, subsets of structurally similar compounds with different mechanisms of action can be identified in the graph representation and selected from ligand datasets.

A general feature of all similarity-based compound networks introduced until recently is that they rely on calculated whole-molecule similarity,[8] typically Tanimoto similarity[14] of chosen molecular representations (such as fingerprints). While the calculation of Tanimoto similarity is usually appropriate for chemoinformatics tasks including network analysis, its use is often insufficient for medicinal chemistry applications. This is the case because in medicinal chemistry, a major focus is the identification of regions in molecules and structural modifications that determine observed SAR characteristics of a compound series or a specific mechanism of action. These "local" and pharmacophore-type analyses are difficult to carry out on the basis of calculated whole-molecule similarity relationships and typically require subsequent structural comparisons.[8] For example, if compounds share 75% global structural similarity,

they still need to be individually compared in order to deduce structural changes that might determine their SAR and/or mechanistic features. This general shortcoming of whole-molecule similarity assessment also affects mechanistic comparisons on the basis of NSGs.[13] In this case, similar ligands with different mechanisms of action must be selected from the network representation and their structures must subsequently be analyzed in order to understand modifications that lead to mechanism hopping. Importantly, the network representation itself does not reveal such changes.

In order to address limitations in structural interpretability associated with calculated whole-molecule similarity values, a molecular network structure termed bipartite matching molecular series graph (BMMSG) has recently been introduced.[15] In this case, calculated similarity values are replaced with defined substructure relationships. This is facilitated by systematically determining all possible substructure relationships between dataset compounds utilizing the matched molecular pair (MMP) formalism.[16] An MMP is defined as a pair of compounds that only differ by the exchange of a single substructure (or, in other words, by a single chemical transformation), for example, a specific R-group or ring.

In BMMSGs, structural relationships between MMP compounds are also indicated by edges, in analogy to calculated pair-wise similarity values utilized for other network representations. Compared to conventional similarity-based networks, a major advance provided by the BMMSG data structure is that this network immediately reveals structural modifications that distinguish compounds from each other. This enhanced chemical interpretability is also the major attraction of the MMP-based compound network for medicinal chemistry applications.

Because a major goal of mechanism of action studies in medicinal chemistry is the identification of structural features of active compounds that are responsible for a specific mechanism, we introduce an extension of the BMMSG structure that incorporates mechanism of action information and makes it readily possible to determine structural changes implicated in mechanism hopping. Structural subset hierarchies derived from the graph representation provide high-resolution views of structural modifications of compound frameworks that lead to different mechanisms of action in pairs of molecules and analog series.

## Methods

### Matched molecular pairs

For dataset compounds, MMPs were systematically generated using an in-house implementation of the algorithm of Hussain and Rea.[17] Following this approach, single bonds that are not part of a ring system are systematically deleted. The process yields a table of fragment pairs. Fragments shared by molecule pairs are termed keys while distinguishing substructures are recorded as values for these keys. If one single bond is deleted ("single cut"), a compound yields two fragments. Each of these fragments is then once indexed as a key and the other as the associated value. If two single bonds ("double cut") or three single bonds ("triple cut") are deleted, a core fragment and two or three substituents are produced, respectively. These substituents are collectively stored as the key and the core fragment as the corresponding value. All MMPs are identified from the index by searching for keys having more than one value. In our implementation, compound pairs only qualify as MMPs if values do not contain more heavy atoms than keys. However, this restriction is not an essential feature and can easily be omitted if compounds with small core structures (*e.g.*, a single ring) are under investigation. Routines to generate and display MMPs were implemented in Java using the OpenEye chemistry toolkit.[18]

### Bipartite graph representation

The MMP table is utilized as a source for the generation of a bipartite graph consisting of two types of nodes, *i.e.*, key nodes and molecule nodes.[15] Keys are displayed if they are associated with more than one value. A key node is connected through edges to all nodes of molecules that contain this key fragment. Accordingly, edges correspond to values and are graphically associated with value fragments.

A matching molecular series (MMS) is defined as a series of compounds that only differ by a single structural change at a specific site. Accordingly, each key node represents an MMS, which can overlap because a compound that differs at one site from a subset of molecules might differ at another site from another subset. All molecules of a series that are only connected to a single key node can be represented as a "super node". It displays the key node as a rectangle that contains all compound nodes shown as individual squares. If a key describes a subset of a series represented by another, it is omitted from the graph to reduce visual complexity. However, subset relationships between all keys are displayed in a separate subset hierarchy (as further detailed in the Results and discussion section). Because the BMMSG structure captures all possible MMS in a dataset, its edges (values) comprehensively represent all substructural relationships between dataset compounds. In the original BMMSG implementation, all molecule nodes were colored according to compound potency to enable SAR analysis.[15]

### Molecular mechanism-based BMMSG

In our receptor ligand analysis, four different mechanisms of action were considered: (1) *agonist* (activates a receptor through

**Table 1** Receptor ligand sets[a]

| Target | Class | Ligand mechanism | No. of compounds | Potency (Ki) Min/μM | Max/nM |
|---|---|---|---|---|---|
| Adenosine A1 receptor | AA1 | Agonist | 107 | 12 | 0.2 |
| | | Partial agonist | 54 | 5 | 2 |
| | | Antagonist | 52 | 66 | 0.4 |
| | | Inverse agonist | 94 | 1 | 0.3 |
| Muscarinic acetylcholine receptor M1 | AM1 | Agonist | 26 | 282 | 2.5 |
| | | Partial agonist | 49 | 29 | 0.3 |
| | | Antagonist | 73 | 30 | 0.1 |
| Histamine H3 receptor | H3R | Agonist | 44 | 11 | 0.3 |
| | | Partial agonist | 46 | 10 | 0.9 |
| | | Antagonist | 31 | 2 | 0.08 |
| | | Inverse agonist | 92 | 0.2 | 0.1 |

[a] For three receptor ligand datasets, the target name, class abbreviation, mechanisms of action, compound numbers, and minimum and maximum potency values are reported.

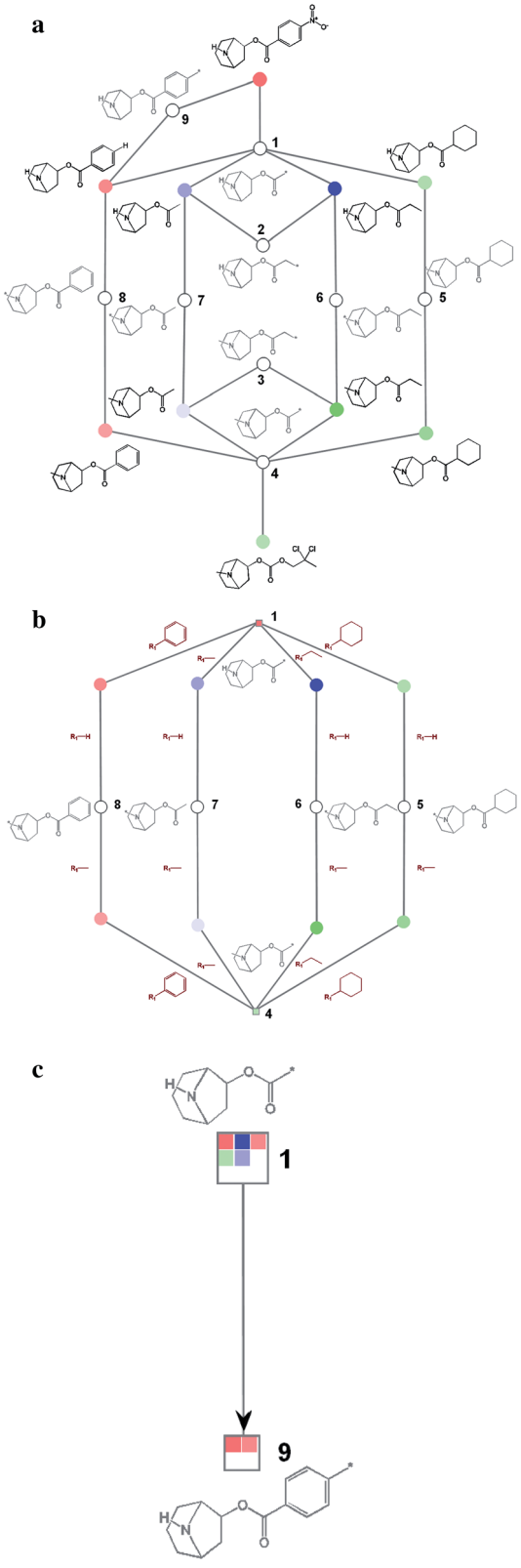



**Fig. 1** Prototypic graph. A schematic representation of the mechanism-based BMMSG structure is shown. (a) The graph for an exemplary dataset that contains all possible key nodes is displayed. These key nodes (numbered 1–9) are colored white whereas all molecule nodes are colored according to their mechanisms of action (agonists: blue, partial agonists: green, antagonists: red, and inverse agonists: gray) and shaded according to the compound potency. The structures of corresponding molecules and keys (shared fragments) are shown in black and gray, respectively. Substituent positions in key fragments are indicated by asterisks. (b) The reduced graph is shown after removal of key nodes involved in subset relationships. Keys 1 and 4 are now represented as (single compound-containing) super nodes. In addition, values associated with keys, *i.e.*, the structures of substituents, are displayed (in brown) next to the edges representing them. (c) The hierarchical subset relationship between keys 1 and 9 is displayed. Here, all compounds associated with these keys are contained in super nodes.

binding to its primary ligand binding site), (2) *partial agonist* (also competes with the physiological receptor ligand but does not exhibit full activation potential), (3) *antagonist* (prevents receptor activation by blocking the primary ligand binding site), and (4) *inverse agonist* (stabilizes an inactive conformation of the receptor and thus prevents activation).

In order to incorporate mechanism of action information into BMMSG representations, we introduced a mechanism-based node coloring scheme as previously suggested for NSGs.[13] Nodes representing agonists were colored blue, nodes representing partial agonists green, antagonists red, and inverse agonists gray. Furthermore, compound potency information was accounted for by shading such that darker shades indicate increasing and lighter shades decreasing potency values (with the degree of shading scaled according to the overall potency range within each subset of ligands having a specific mechanism). In contrast to molecule nodes, all key nodes were colored white. These graph representations were implemented using the Java package JUNG.[19] The graph layout was generated separately for each subgraph (see Results and discussion) using a JUNG implementation of an algorithm to display self-organizing maps.

### Datasets

Compounds active against three G protein coupled receptors (GPCRs) with different mechanisms of action were assembled from ChEMBL.[20] Mechanistic annotations for these receptor ligands were taken from the original publications. Molecules were classified as agonists, partial agonists, antagonists, and inverse agonists, provided the mechanism was uniquely defined. In the case of multiple annotations, the following rules were applied. If agonist and partial agonist activities were reported, ligands were classified as partial agonists. If antagonist and inverse agonist activities were reported, ligands were classified as inverse agonists (because apparent antagonists are often found to be inverse agonists[3]). To ensure consistency in the use of potency measurements, only Ki values were considered as potency annotations. For compounds with multiple Ki measurements, the geometric mean was calculated. The composition of the receptor ligand datasets is summarized in Table 1.

## Results and discussion

### Exemplary graph representation

In Fig. 1, the design elements of the mechanism-based BMMSG (M-BMMSG) are illustrated. Fig. 1a shows a graphical representation with all key nodes (white) and molecule nodes (with mechanism-based coloring) for a model dataset. Each key node

represents the common substructure (core) of all molecule nodes connected to it. In key nodes, attachment points for substituents are labeled with asterisks (the minimal structural modification distinguishing a key and a molecule is the hydrogen substituent). The set of molecules attached to each key node represents an MMS. The graph also illustrates that individual molecules might belong to different series based on their substitution patterns. Fig. 1b shows the M-BMMSG following removal of three key nodes (2, 3, and 9) that are involved in subset relationships and introduction of super nodes. This reduced graph is utilized as the standard representation for the display of compound datasets. The structures of substituents that distinguish keys and molecules from each other are shown next to their edges. In Fig. 1c, an exemplary subset hierarchy involving keys 1 and 9 is shown. Key 9 was removed from the graph because the series it represents is a subset of the one represented by key 1. As further discussed in the following, subset hierarchies are displayed in separate tree structures that complement the M-BMMSG representation and help to elucidate structural changes implicated in mechanism hops.

### Dataset representations

Fig. 2 shows the final (reduced) M-BMMSGs for the three different receptor ligand sets we analyzed. The graphs of the complete datasets consist of different disjoint subgraphs. Compounds in each subgraph do not form substructure relationships with compounds in other subgraphs. The topology of the three M-BMMSGs in Fig. 2 notably differs, which indicates the presence of different structural relationships and different degrees of structural diversity among the ligands in these sets. The AA1 set in Fig. 2a consists of 307 ligands and its M-BMMSG contains a large subgraph that predominantly consists of agonists (blue) and partial agonists (green) with varying potencies. In addition, there is a medium-sized subgraph consisting of a region with many highly potent inverse agonists (dark gray) and another region with agonists and partial agonists along with a few antagonists. These regions are only connected through a single edge and key node. The AA1 graph also contains three small subgraphs that consist of highly or weakly potent antagonists (red) and six individual compound series, each represented by a super node, which form no structural relationships to others. Thus, overall there is a clear separation of compounds by mechanism of action across these subgraphs, with many structural relationships formed between compounds sharing the same mechanism. Similar observations are made for the M-BMMSG of the AM1 set in Fig. 2b that is approximately half the size of the AA1 set (148 compounds). However, in this case, six subgraphs of comparable size are formed and 11 individual compound series, thus indicating an overall higher degree of structural diversity among these receptor ligands. Moreover, as further discussed below, three of the six subgraphs consist of compounds having three different mechanisms of action. In contrast to the AM1 set, the M-BMMSG of the H3R set in Fig. 2c (with 213 compounds) consists of a major and in part densely connected graph component and five individual series, thus indicating the presence of many substructure relationships between these ligands (corresponding to a higher degree of structural homogeneity than in the previous cases). In the major graph

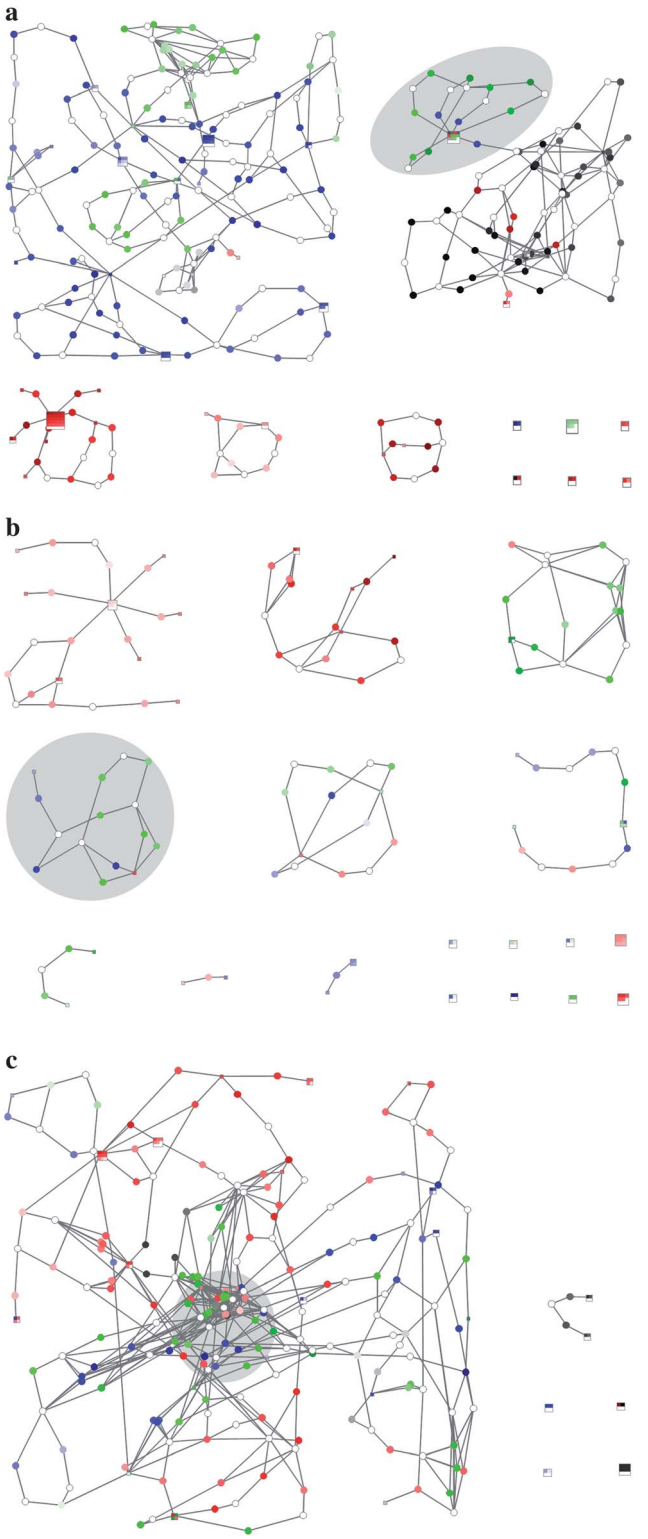



**Fig. 2** Mechanism-based BMMSGs. Shown are M-BMMSG representations for (a) AA1 (consisting of 11 separate subgraphs), (b) AM1 (17 subgraphs) and (c) H3R (6 subgraphs). Selected regions containing ligands with multiple mechanisms of action are displayed on a gray background. Compound nodes are colored according to Fig. 1 and white nodes are key nodes. For clarity, selected key nodes are displayed as super nodes.

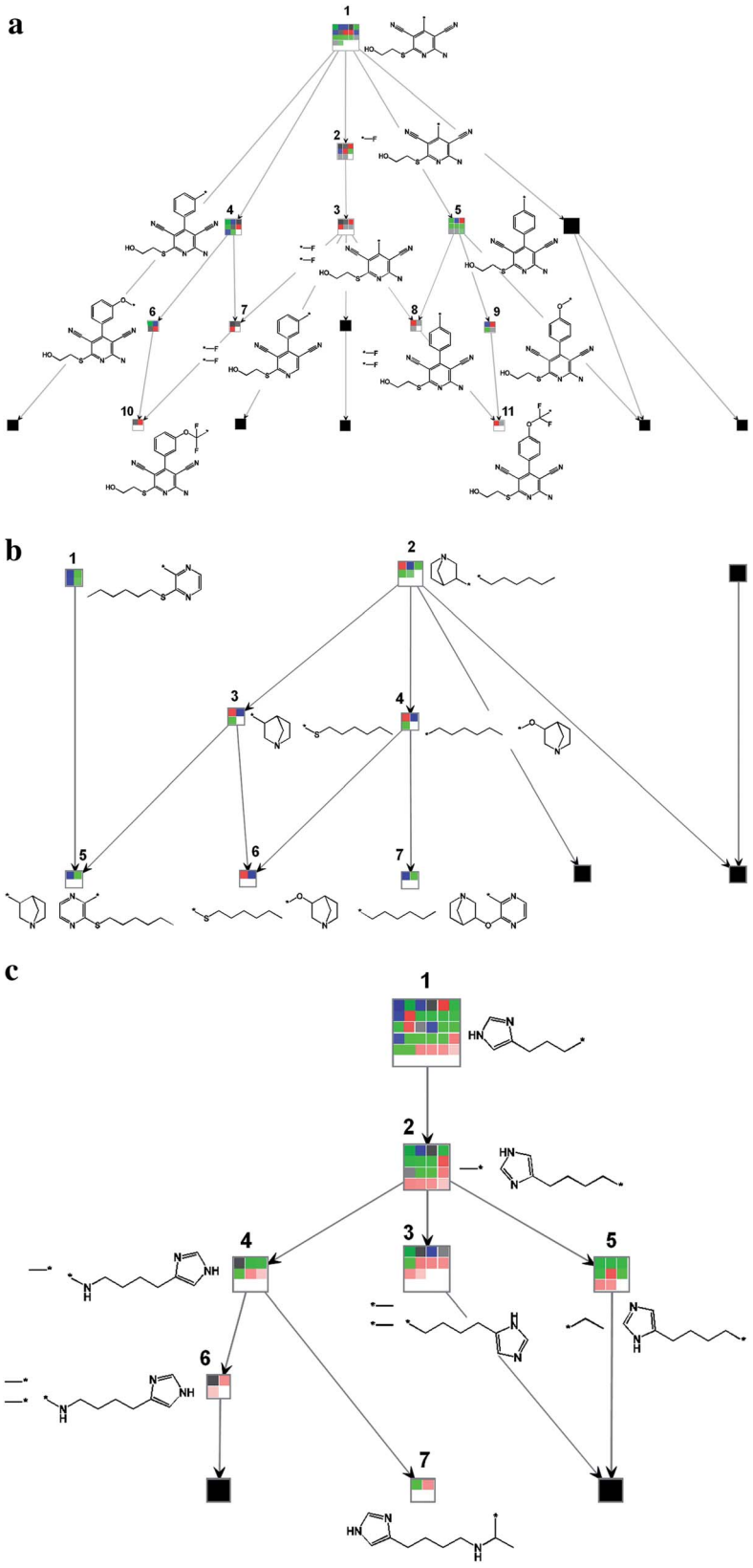



**Fig. 3** Key subset relationships. Hierarchical subset relationships between keys are shown for the mechanism hopping regions highlighted in Fig. 2. Super nodes representing keys associated with compounds having the same mechanism of action are shown in black. For the remaining super nodes, the shared substructure (core) is displayed and molecule nodes are color-coded according to their mechanisms. Nodes containing mechanism hops are numbered. (a) AA1, (b) AM1, and (c) H3R.

component, there are regions predominantly populated with antagonists. However, the most densely connected region in the center of this large subgraph contains compounds with all four mechanisms including many partial agonists.

### Mechanism hopping regions

The clustering of structurally related compounds by mechanism that is observed at varying degrees in M-BMMSGs in Fig. 2 is a characteristic feature of these graphical representations. However, it is not the only one; rather, all three M-BMMSGs also contain subgraphs, or parts of subgraphs, that are composed of structurally related yet mechanistically distinct ligands (indicated by different node colors), which are easily identified through visual inspection of the graph representations. These regions are prime candidates for the exploration of mechanism hops. In each of the three M-BMMSGs in Fig. 2, one exemplary mechanism hopping region is highlighted. These regions are further analyzed in the following.

### Critical structural modifications

Mechanism hopping regions are explored in order to identify structural modifications of ligands that lead to mechanistic changes. In Tanimoto similarity-based graph representations such as M-NSGs,[13] mechanism hopping regions can also be identified. However, compound subsets comprising such regions must then be selected and separately analyzed. By contrast, for the identification of critical structural modifications leading to mechanism hops, the M-BMMSG structure with its intrinsic subset hierarchy is ideally suited. In Fig. 3, the complete subset hierarchies for the mechanism hopping regions highlighted in Fig. 2 are displayed. These subset hierarchies contain all keys that are relevant for mechanism hopping as super nodes including those that are omitted from the final M-BMMSG representations (but are a part of the underlying graph structure). Each super node within a hierarchy is associated with the shared substructure and the molecules it contains are associated with the distinguishing structural fragments (substituents). As shown in Fig. 3, following the subset hierarchies from the top to the bottom, super nodes are associated with key structures of increasing size and continuously smaller ligand subsets. Importantly, within these compound subsets, there is increasing separation of ligands by mechanism along the tree structure, as also illustrated in Fig. 3. Nodes that exclusively contain compounds sharing the same mechanism are colored black because they are not relevant for mechanism hopping analysis. However, all other super nodes within the hierarchy contain mechanism hops and nodes at the bottom of the tree represent pairs of ligands with individual mechanism hops. From these super nodes, substituents that lead to mechanistic changes can be directly selected, as illustrated in Fig. 4. This figure shows the substituents of compounds from all super nodes containing mechanism hops. Each super node is associated with a key structure having a defined substitution site (if the corresponding MMPs result from a single cut; see Methods) or two or three sites (if the corresponding MMPs result from double or triple cuts, respectively, and the resulting keys consist of two or three fragments). Hence, in addition to conventional R-groups, central structural moieties

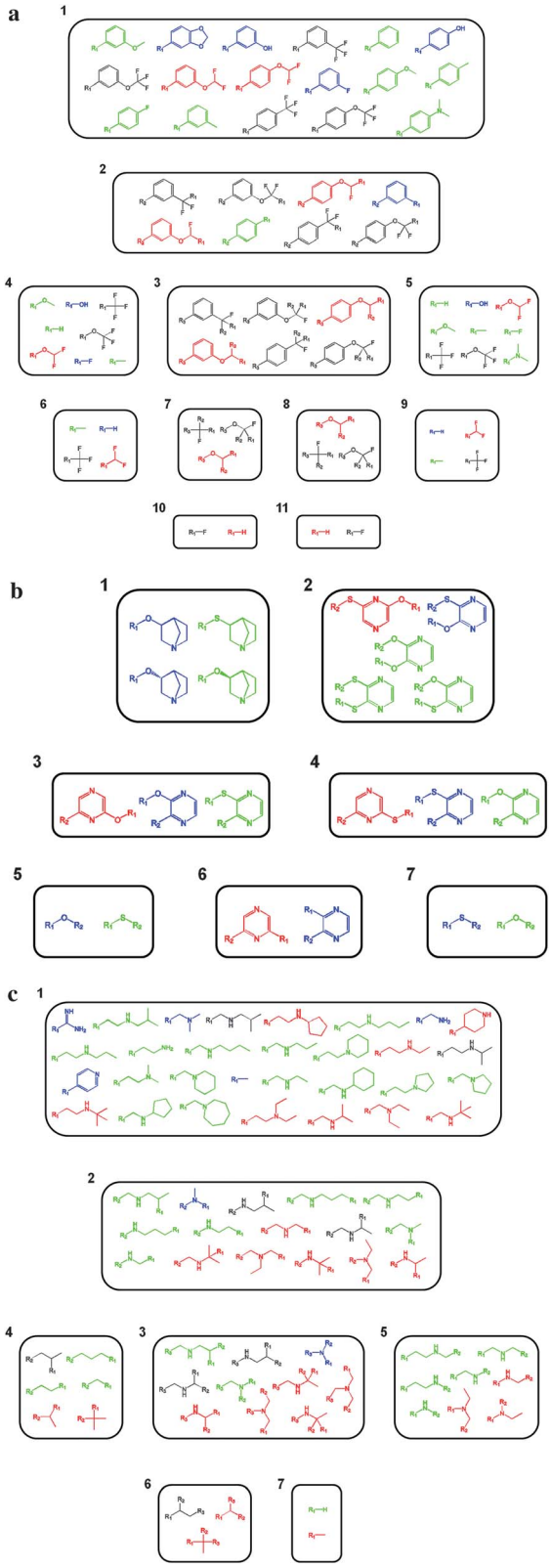



**Fig. 4** Value fragments associated with mechanism hops. For super nodes numbered according to Fig. 3, substituents are shown that lead to mechanism hopping (and are color-coded accordingly). (a) AA1, (b) AM1, and (c) H3R.

can also be replaced in ligands with different mechanisms. This spectrum of possible structural transformations is illustrated in Fig. 4a showing substituents of the subset hierarchy of AA1 ligands according to Fig. 3a. For example, in node 1, all substituents contain, with one exception, a conserved phenyl moiety with a single R-group at the ring. The presence of different substituents in the *meta* or *para* position at these phenyl rings characterizes ligands with different mechanisms. In node 2 in Fig. 4a, substructures with two substitution sites are exchanged (because the corresponding key consists of two fragments; see Fig. 3a). Smaller ligand and corresponding substituent sets at subsequent nodes reveal increasingly specific structural modifications. For example, in node 5, hydrogen, methyl, fluoro, secondary amino, or methoxy groups at the single substitution site are characteristic of partial AA1 agonists, whereas di- or tri-fluoro substituents are found in inverse agonists or antagonists. At the bottom, substituent pairs in nodes 6 and 7 reveal that the introduction of a fluorine at the single site of the largest key structures (effectively yielding a tri-fluoro-ether or -methyl group) converts AA1 antagonists within this series into inverse agonists.

In Fig. 4b, substituent sets for the nodes of the AM1 hierarchy in Fig. 3b also reveal well-defined replacements leading to mechanistic changes. For example, in nodes 3 and 4, exchanges of differently substituted pyrazine rings convert partial agonists or agonists into antagonists. Moreover, the substituent pair in node 6 (one of the terminal nodes of the hierarchy) reveals that a change in the relative position of the pyrazine ring in this series of compounds is sufficient to cause an agonist-to-antagonist switch. Equivalent observations are also made for the H3R substituent sets in Fig. 4c. For example, the substituent pair in the terminal node 7 indicates that the introduction of a methyl group (replacing a hydrogen) at the single substituent site of the corresponding key structure is sufficient to convert a partial H3R agonist into an antagonist. Thus, taken together, these findings illustrate that subset hierarchies of mechanism hopping regions in M-BMMSGs reveal structural modifications that cause mechanism hops.

### Mechanism hopping in analog series

These structural transformations can also be directly traced back to the compounds from which they originate. Accordingly, a series of ligands with multiple mechanism hops can be selected from M-BMMSGs, which is particularly attractive for medicinal chemistry applications. Fig. 5 shows examples of analog series taken from mechanism hopping regions where substitutions at a single site change the mechanism of action in different ways. In Fig. 5a, a series of AA1 ligands is shown. Here, the presence of a hydroxyl group at the *para* position of the phenyl moiety characterizes an agonist. Replacement of this hydroxyl group with a methoxy group leads to a partial agonist. However, the introduction of a di- or tri-fluoro-methoxy group at this position converts the partial agonist into an antagonist or inverse agonist, respectively. In the series of AM1 ligands in Fig. 5b, changing the 1,2-substitution pattern at the pyrazine ring to a 1,3-substitution pattern (*i.e.*, facilitating a positional shift) leads to an elongated analog structure and converts the agonist and partial agonist into an antagonist. Finally, in the series of H3R ligands in Fig. 5c, extending the

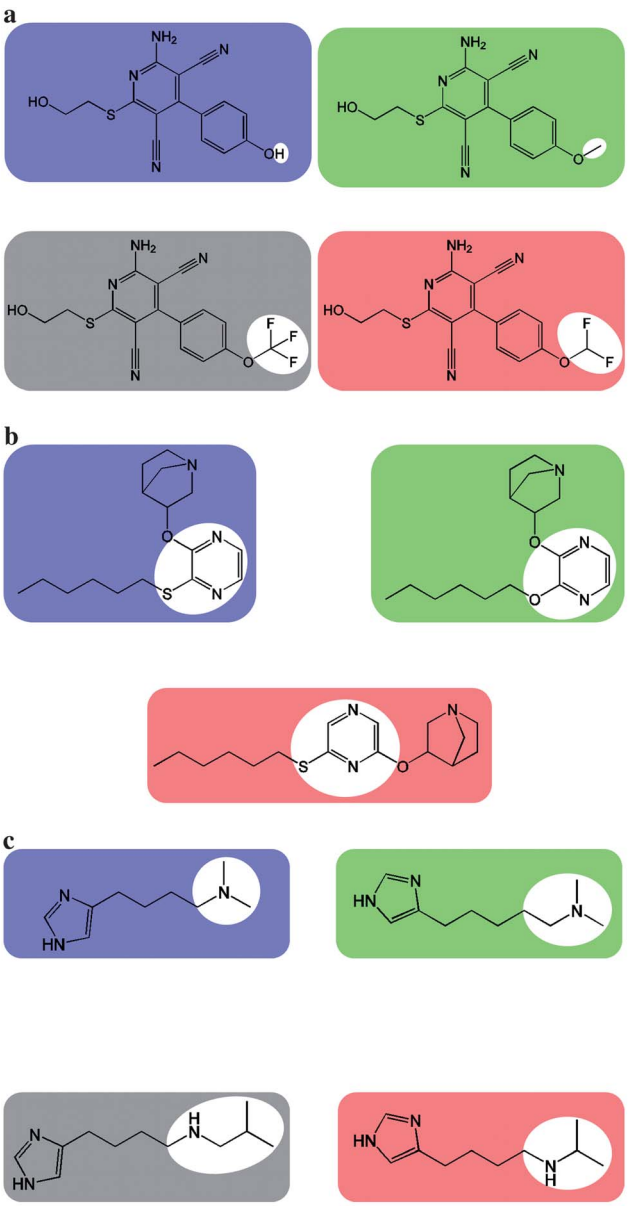



**Fig. 5** Selected compounds. For each dataset, exemplary compounds are shown that are distinguished by single chemical transformations and have different mechanisms of action (indicated by the mechanism-based color code). (a) AA1, (b) AM1, and (c) H3R.

linker length of the di-methyl-amine substituent transforms an agonist into a partial agonist. However, replacing the di-methyl-amine group in the agonist with an isopropyl- or isobutyl-amine converts it into an antagonist or inverse agonist, respectively. Thus, the hierarchical analysis of structural transformations that occur in mechanism hopping regions identified in M-BMMSGs also makes it possible to select a series of analogs with defined structural modifications that lead to mechanistic changes.

## Conclusions

Herein we have introduced a graphical method to study the structural basis of mechanism hopping in sets of receptor ligands.

Central features of this approach include the ability to identify local mechanism hopping regions in M-BMMSGs and directly select structural modifications that are responsible for mechanistic changes from corresponding key node subset hierarchies. These subset hierarchies provide high-resolution views of mechanism hops and underlying structural changes because there often is an increasing separation of ligands by mechanism along the tree structure. At each level of the tree, the corresponding structural modifications are immediately accessible. In practical applications, one initially generates an M-BMMSG representation of an entire dataset and then selects mechanism hopping regions, if available, for subset hierarchy display and the identification of important structural modifications. Subset hierarchies of mechanism hopping regions often describe series of ligands with multiple mechanism hops. The M-BMMSG representations of the original datasets then also offer the opportunity to further analyze the neighborhood of such series and their structural organization. Thus, the M-BMMSG approach facilitates the analysis of mechanism hopping at multiple levels.

## References


1 T. Kenakin, *Trends Pharmacol. Sci.*, 2004, **25**, 186–192.
2 B. T. Zhu, *Biomed. Pharmacol.*, 2005, **59**, 76–89.
3 P. J. Greasley and J. C. Clapham, *Eur. J. Pharmacol.*, 2006, **553**, 1–9.
4 J. Kraut, *Science*, 1988, **242**, 533–540.
5 T. C. Bruice and S. J. Benkovic, *Biochemistry*, 2000, **39**, 6267–6274.
6 N. M. Goodey and S. J. Benkovic, *Nat. Chem. Biol.*, 2008, **4**, 474–482.
7 L. Peltason and J. Bajorath, *Future Med. Chem.*, 2009, **1**, 451–466.
8 A. M. Wassermann, M. Wawer and J. Bajorath, *J. Med. Chem.*, 2010, **53**, 8209–8223.
9 M. Wawer, E. Lounkine, A. M. Wassermann and J. Bajorath, *Drug Discovery Today*, 2010, **15**, 630–639.
10 A. L. Hopkins, *Nat. Chem. Biol.*, 2008, **4**, 682–690.
11 M. J. Keiser, J. J. Irwin and B. K. Shoichet, *Biochemistry*, 2010, **49**, 10267–10276.
12 M. Wawer, L. Peltason, N. Weskamp, A. Teckentrup and J. Bajorath, *J. Med. Chem.*, 2008, **51**, 6075–6084.
13 P. Iyer, D. Stumpfe and J. Bajorath, *J. Chem. Inf. Model.*, 2011, **51**, 1281–1286.
14 P. Willett, *J. Med. Chem.*, 2005, **48**, 1–17.
15 M. Wawer and J. Bajorath, *J. Med. Chem.*, 2011, **54**, 2944–2951.
16 P. W. Kenny and J. Sadowski, in *Chemoinformatics in Drug Discovery*, ed. T. I. Oprea, Wiley-VCH, Weinheim, Germany, 2004, pp. 271–285.
17 J. Hussain and C. Rea, *J. Chem. Inf. Model.*, 2010, **50**, 339–348.
18 *OEChem TK version 1.7.4.3*, OpenEye Scientific Software Inc., Santa Fe, New Mexico, 2010.
19 *Java Universal Network/Graph Framework version 2.0.1.*, http://jung.sourceforge.net/, accessed 11 January, 2010.
20 *ChEMBL*, http://www.ebi.ac.uk/chembl/, accessed March 2, 2011.

# Summary

This chapter reports the modification of an existing graphical landscape representation that integrates systematic SAR information to include the mechanism of action of receptor ligands. This data structure organizes compounds using clearly defined substructure relationships that facilitates a smooth transition from SAR and mechanistic explorations to the identification of structural modifications associated with mechanism hops. In addition, subset hierarchies allow inspection of these chemical changes at various levels of resolution ranging from analog series to compound pairs. Separation of ligands with respect to their mechanism of action increases with increase in the hierarchical levels. Information obtained after exploration of substitutions that bring about mechanistic transitions can be utilized in various compound design and optimization applications.

A common feature of the approaches reported thus far is that these have been designed to focus on compound activities against a single target. However, methodologies that can be applied to compound sets with activities against multiple targets are also of interest, especially in the study of polypharmacology. A novel multi-target graphical representation is reported in the following chapter.

# Chapter 6

# Representation of multi-target activity landscapes through target pair-based compound encoding in self-organizing maps

## Introduction

Systematic SAR investigations are often performed using activity landscape modeling as it provides a framework for the integration of structural similarity relationships with potency information. Traditional SAR analyses focus on compounds with activity annotations against individual targets. Attempts have been made to design selectivity landscape representations to investigate potency distributions for two targets. However, activity landscapes that consistently account for pair-wise compound similarities and multi-target activities are necessary to study polypharmacological compounds, i.e. compounds that interact with different targets. For this purpose, a novel multi-target activity landscape has been introduced where the chemical similarities were projected into 2D space using self-organizing maps (SOMs) [1]. Compounds with reported activities against multiple targets were represented as arrays of cells colored according to binned pairwise target potency differences. Using this landscape model, it was possible to rationalize discontinuity in multi-target ac-

tivity space. This multi-target landscape representation has successfully been applied to ligand sets with activities against three to five targets.

[1] Kohonen T. *Self-organizing maps*, Springer, Heidelberg, Germany, **1996**.

*Chem Biol Drug Des 2011;* **78:** 778–786

Research Article



doi: 10.1111/j.1747-0285.2011.01235.x

# Representation of Multi-Target Activity Landscapes Through Target Pair-Based Compound Encoding in Self-Organizing Maps

**Preeti Iyer 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*

**Activity landscape representations provide access to structure-activity relationships information in compound data sets. In general, activity landscape models integrate molecular similarity relationships with biological activity data. Typically, activity against a single target is monitored. However, for steadily increasing numbers of compounds, activity against multiple targets is reported, resulting in an opportunity, and often a need, to explore multi-target structure-activity relationships. It would be attractive to utilize activity landscape representations to aid in this process, but the design of activity landscapes for multiple targets is a complicated task. Only recently has a first multi-target landscape model been introduced, consisting of an annotated compound network focused on the systematic detection of activity cliffs. Herein, we report a conceptually different multi-target activity landscape design that is based on a 2D projection of chemical reference space using self-organizing maps and encodes compounds as arrays of pair-wise target activity relationships. In this context, we introduce the concept of discontinuity in multi-target activity space. The well-ordered activity landscape model highlights centers of discontinuity in activity space and is straightforward to interpret. It has been applied to analyze compound data sets with three, four, and five target annotations and identify multi-target structure-activity relationships determinants in analog series.**

**Key words:** activity landscapes, data mining, graphical representations, multi-target SARs, self-organizing maps, structure-activity relationships

Received July 25, 2011, revised August 24, 2011, and accepted for publication August 27, 2011

Graphical representations of activity landscapes have become increasingly popular for the qualitative analysis of structure-activity relationships (SARs), identification of activity cliffs, characterization of local SAR environments, and extraction of SAR information from large compound data sets (1–3). In general, an activity landscape is best rationalized as a hypersurface in chemical space that accounts for the biological activity of a compound set. If one envisions a 2D (x, y-) projection of chemical space in which distances between compounds indicate molecular (dis-)similarity relationships, compound potency can be added as the third (z-) dimension to this representation. Then, a potency surface can be interpolated, giving rise to intuitive 3D landscape representations (1,3) that are reminiscent of geographical maps. Such 3D landscapes have often been discussed in idealized form (3) but can also be approximated for actual compound data sets (4). In addition to 3D activity landscapes, a variety of different 2D landscape representations have been generated (3,5–7). In fact, the first landscape views of compound data sets were 2D representations, beginning with the introduction of so-called structure-activity similarity maps (7), for which a number of derivatives have been reported (8,9). In addition to such graphs, which systematically account for similarity and potency relationships on the basis of pair-wise compound comparisons, annotated molecular network representations have become popular for 2D landscape design (5,6).

Regardless of their dimensionality or design characteristics, conventional activity landscapes monitor the activity of compound data sets against individual targets (3). However, with mounting evidence of polypharmacological compound behavior and network pharmacology (10–12), multi-target compound activity profiles are increasingly considered in SAR analysis (13–15). Ideally, one would also like to employ activity landscape models for the evaluation of multi-target SARs. An extension of conventional activity landscape models has been introduced where network-like similarity graphs, a molecular network–based 2D landscape representation (5), have been transformed into dual-target selectivity landscapes by considering potency ratios instead of single-target potency values (16). On this basis, the concept of selectivity cliffs has been introduced (16). However, the design of activity landscape representations for three or more targets has proven to be a difficult task. Recently, a first design of a multi-target activity landscape (MTAL) has been introduced (17), which captures compound potency relationships across multiple targets in a formally consistent manner by introducing numerical codes for multi-target activity profiles. These codes were mapped onto compounds (nodes) in similarity-based compound data set layouts where edges indicate the presence of single- or multi-target activity cliffs formed by pairs of compounds (17).

We have continued to explore alternative ways to generate MTAL models and present herein a conceptually different approach, which

departs from molecular network representations. A neural network–based projection of chemical reference space is utilized to provide a similarity-based 2D organization of compounds with multiple target annotations. Then, compounds are represented as arrays of pair-wise target potency relationships, which add a new data structure to the self-organizing map (SOM)-based structural organization. Elements of the arrays are color-coded according to the degree of activity discontinuity a compound displays. The resulting data structure is straightforward to analyze and reveals multi-target SAR hotspots.

## Methods and Materials

### *Data sets*

Compound data sets with multiple target annotations were extracted from ChEMBL[a] (18). Only potency measurements ($K_i$ or, if not available, $IC_{50}$ values) with defined endpoints and designated maximum confidence level were considered. Other less well-defined potency measures such as '% inhibition' or reported ranges of Ki or $IC_{50}$ values were not considered because of their intrinsic accuracy limitations. Data sets with reported activity against three, four, and five targets were obtained, as summarized in Table 1. These sets include 342 adenosine receptor antagonists with reported activity against three different adenosine receptors (designated 3AR), 98 opioid-like receptor antagonists with activity against three different opioid receptors and the nociceptin receptor (4OR), and 53 inhibitors of five thrombin-related serine proteases (5PR). In currently available public domain compound data, it is difficult to find significant numbers of compounds that share five or more target annotations (17). Therefore, the compound sets studied herein reflect the current spectrum of multi-target activity data.

### *Molecular representation*

For all compounds, the stereochemistry-sensitive form of the Extended Connectivity Fingerprint with bond diameter 4 (ECFP4)[b] (19) was calculated as an input representation for compound classification. ECFP4 monitors layered atom environments in test compounds and is generally considered a high-resolution fingerprint.

### *Chemical reference space projection*

Pair-wise Tanimoto coefficient (Tc) values were calculated for all compound fingerprints to constitute a low-complexity co-ordinate-free reference space. A 2D projection of this chemical reference space was obtained through the calculation of a SOM (20) using the SONNIA program[c] (20). Self-organizing map is a neural network method that facilitates dimension reduction in chemical space representations and assigns compounds to neurons organized in a plane. This 2D projection mirrors similarity relationships by clustering compounds such that similar ones are assigned to the same or adjacent neurons. Hence, distance between compounds in a SOM is a measure of (dis-)similarity (i.e., the shorter the distance between compounds is, the more similar they are). Pair-wise Tc values were used as input for SOM calculation. The dimension for neuron generation was set to the size of each data set. SONNIA default parameters were applied to derive the neuron grid.

### *Multi-target activity landscape*

The grid of neurons of each SOM was used as the template for activity landscape design. Neurons to which compounds were assigned were shown with black borders and 'empty' neurons with gray borders. Compounds were represented as follows: (i) For each compound, all possible pairs of targets were enumerated (i.e., for activity against three targets 1, 2, and 3, three pairs 1_2, 1_3, and 2_3 were obtained). (ii) Each compound was represented by *n* squares, each accounting for a target pair, drawn on a light blue background (to distinguish different compounds from each other). Each square was labeled with the corresponding target pair (e.g., 1_2) (iii) For each pair, the logarithmic potency difference in the compound was calculated. Potency differences were assigned to five bins, and the squares were color-coded according to the magnitude of the difference: $\Delta$pKi (or $\Delta IC_{50}$) $< 1$ (green), $1 \leq \Delta pKi < 2$ (light green), $2 \leq \Delta pKi < 3$ (yellow), $3 \leq \Delta pKi < 4$ (orange), and $\Delta pKi \geq 4$ (red). Thus, squares with smallest target pair-wise potency differences within one order of magnitude were colored green, and squares with largest potency differences in four or more orders of magnitude were colored red. For interactive display and navigation, the activity landscape design was implemented in Java.

## Results and Discussion

The analysis of multi-target SARs generally is a complicated task. Here, we present a model of a MTAL that is designed to provide an intuitive access to local multi-target SAR components with a particular focus on regions that are characterized by high discontinuity in multi-target activity space. However, as shown in the following, regions of multi-target SAR continuity can also be identified in these representations. In MTALs, regions of continuity correspond to compounds having similar activity against each target. Hence, such regions are less informative for the analysis of multi-target SARs than regions of significant discontinuity. This is the case because discontinuity is introduced by compounds that display differential activity against multiple targets, and from such compounds, SAR determinants might be deduced.

**Table 1:** Compound data sets with multi-target activity

| Class | Activity | Size | No of targets | Targets |
|---|---|---|---|---|
| 3AR | Adenosine receptor antagonists | 342 | 3 | Adenosine receptors A1, A2a, A3 |
| 4OR | Opioid-like receptor antagonists | 98 | 4 | $\delta$-, $\kappa$-, $\mu$-Opioid receptors, Nociceptin receptor |
| 5PR | Urokinase-type plasminogen activator-like inhibitors | 53 | 5 | Thrombin, Plasminogen, Coagulation factor X, Urokinase-type plasminogen activator, Matriptase |

For the three data sets assembled from ChEMBL, the class abbreviation, activity, size, and target information are provided.

### *Methodological concept and design elements*

The generation of activity landscape models principally requires the integration of molecular similarity relationships and compound potency information. For activity against individual targets (or at most two targets), potency (selectivity) information can be directly added to chemical space representations or systematically monitored pair-wise compound similarity relationships. For activity against three or more targets, other approaches are required. For MTAL design, the chemical space/similarity relationship display is analogous to single-target landscapes (i.e., compound similarity relationships are independent of the number of target annotations). By contrast, a key question is how to best represent compounds in multi-target activity space. For the MTAL model reported herein, we utilize a SOM-based 2D projection of a co-ordinate-free fingerprint space. By computing SOMs, structurally similar compounds are assigned to the same or neighboring neurons resulting in a clustering effect. It should be stressed that the SOM projection provides the initial well-ordered structural reference frame for compound representation. However, to represent a MTAL, a new data structure must be added to it. Hence, to account for multi-target activity, each compound is represented as an array of pair-wise target combinations. Each element (square) of the array is then color-coded according to the potency difference in the compound against the two targets. This design strategy is illustrated in Figure 1. To represent potency differences in a consistent manner throughout the activity landscape, logarithmic potency differences are assigned to five bins. The elements are then color-coded according to the magnitude of the potency difference in a compound against a target pair, from green (potency difference within an order of magnitude) over light green (one to two orders of magnitude), yellow (two to three), and orange (three to four) to red (potency difference in more than four orders of magnitude). As illustrated in Figure 1, this compound representation scheme produces color patterns that reflect differential potency against pairs of targets. In Figure 2A, B, we represent an MTAL model of a largely discontinuous and continuous SAR region, respectively. The region of discontinuity in Figure 2A is characterized by the dominance of mixed color patterns, as discussed in more detail later. Here, it should be noted that the cells in our MTAL model do not convey information about potency directionality (i.e., against which target a compound is highly or weakly potent) because pair-wise potency differences are not directional. However, this information is relevant for the subsequent analysis of strongly discontinuous regions. Therefore, once a discontinuous region of interest has been identified in an MTAL representation, the compound subset forming this region can be displayed together with directional potency differences, as suggested previously for the analysis of multi-target activity cliffs (21). In contrast to discontinuous regions, the region of continuity in Figure 2B is mostly colored in green, reflecting small pair-wise target potency differences. Green regions in MTALs are characteristic of multi-target SAR continuity. To further differentiate between small potency differences resulting from either high or low compound potencies in regions of SAR continuity, the color code can be modified, as illustrated in Figure 2C. For this purpose, absolute potency values are divided into a high ($pKi/IC_{50} > 8$), intermediate ($pKi/IC_{50}$ between 8 and 6), and low ($pKi/IC_{50} < 6$) potency category. Cells representing differences between high, intermediate, and low values are then color-coded in dark blue, blue, and light blue, respectively, which replaces the standard green coloring of cells corresponding to low-potency differences within an order of magnitude. For cells representing potency differences within two orders of magnitude, the standard green coloring scheme still applies, as also illustrated in Figure 2C. However, in our analysis, we predominantly focus on the regions of multi-target SAR discontinuity, as rationalized earlier. Accordingly, the further refined color-coding scheme for continuous MTAL regions is not applied in the examples discussed in the following to limit the complexity of the activity landscape representations.

### *Interpretation of MTAL patterns*

Figure 3A shows the complete MTAL representation for the 4OR data set. The figure illustrates the well-ordered structure of the neuron grid and the compound clustering effect. The grid layout provides an easy access to compound information. The 4OR MTAL reveals different color patterns and a separation of predominantly green compounds from compounds with mixed color patterns. If arrays are colored green, the corresponding compounds display similar potency against multiple targets and hence no apparent selectivity. By contrast, yellow-to-red squares reflect 100- to more than 10 000-fold potency differences against target pairs. As can be seen in Figure 3A, there are no antagonists that consistently display large potency differences against all pairs of targets. Rather, orange and/or red squares usually occur in combination with green squares. This means that a compound displays comparable potency against two or more targets and significantly different potency against at least one or more others. Therefore, such *mixed color*

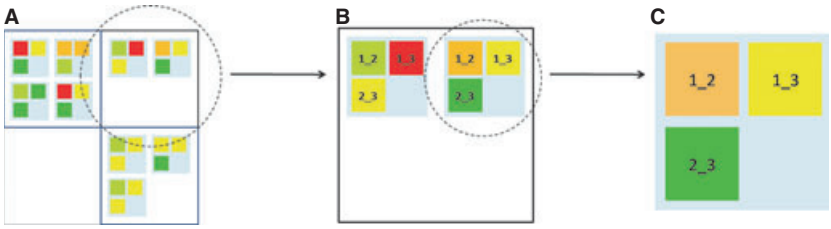



**Figure 1:** Multi-target activity landscape (MTAL) design. (A) A schematic self-organizing map-based MTAL consisting of only four neurons is shown. From the top left in clockwise direction, the neurons contain four, two (encircled), three, and no compounds (with three hypothetical activity annotations). The target pair array representing each compound is depicted on a light blue background. Neurons containing compounds are shown with black borders and empty neurons with gray borders. (B) The neuron containing two compounds encircled in (A) is enlarged (one of the compounds is encircled). Squares are color-coded by binned potency differences and labeled with the corresponding target pairs. (C) The compound encircled in (B) is shown in a close-up view.


**Figure 2:** Exemplary regions of multi-target structure-activity relationships (SAR) discontinuity and continuity. For an exemplary compound data set with activity against three targets, enlarged multi-target activity landscape views of a highly (A) discontinuous and (B) continuous SAR region are shown. In (C), a variant of the standard representation of the continuous region in (B) is shown with modified color code to further differentiate between small potency differences resulting from different levels of compound potency, as explained in the text.


A
B
1_2 1_3 1_4 2_3 2_4 3_4
C
[101]
[102]
[201]
[202]
[203]
[204]
[301]
[302]
[303]

*patterns indicate that a compound introduces a high level of discontinuity in multi-target activity space.*

### Extracting SAR information

What type of SAR information can be obtained from the SOM-based MTAL? A major focal point of activity landscape analysis usually is the identification of prominent activity cliffs in the regions of high local SAR discontinuity (1,3). The multi-target landscape concept previously introduced by Dimova *et al.* (17) specifically aimed at a comprehensive account of single- and multi-target activity cliffs in compound data sets. A multi-target activity cliff is formed by a pair of structurally similar compounds that show significantly different potency against two or more targets. In SAR analysis, one would then compare the structures of these compounds to identify the modification(s) that are responsible for cliff formation. Activity cliffs can also be identified in the newly introduced MTAL design by comparing the colors of corresponding squares in similar compounds. For example, the '2_3' squares of the two adjacent compounds in Figure 1B form a yellow/green combination. Accordingly, one compound has comparable potency against the target pair and the other a 100- to 1000-fold difference in potency. Thus, this compound pair forms an activity cliff. Upon close inspection, a number of such examples become apparent in the landscape view in Figure 3A. However, the detection of activity cliffs is not the primary focal point of our new MTAL method. Rather, its major purpose is the identification of compound subsets (or individual compounds) that are responsible for the introduction of significant discontinuity in multi-target activity space. For example, understanding the basis of such discontinuity is of critical importance for exploring the design of target-selective compounds. In our landscape views, such regions are indicated by mixed color patterns, as illustrated in Figure 1, and become immediately apparent. In Figure 3A, a discontinuous region/compound subset consisting of three adjacent neurons is marked. The rationale for focusing on such regions is as follows: *by comparing similar molecules that do or do not introduce discontinuity in multi-target activity space, substitution sites can be identified that substantially influence or determine multi-target SARs,* as further discussed later.

### Opioid-like receptor antagonists

In Figure 3B, the discontinuous three-neuron region outlined in Figure 3A is depicted, revealing a systematic mixed pattern. The compounds in this region show pronounced potency differences against targets 2 and 4 and very similar potency against targets 1 and 3 and, in part, 1 and 4. Furthermore, significant differences are observed in the behavior of compounds against targets 3 and 4. In Figure 3C, the nine compounds comprising this region are shown (with neuron grid indices defining their origin and position). As can be seen, these compounds form an analog series. In all cases studied here, we have observed that compounds mapping to the same or neighboring neurons were structurally very similar, often series of analogs, which assigns confidence to the SOM-based structural classification underlying the MTAL design. The comparison of the analogs in Figure 3C reveals the presence of two substitution sites at adjacent phenyl moieties (highlighted in Figure 3C) where different R-groups are responsible for the introduction of discontinuity in multi-target activity space. Thus, substitutions at these sites play a major role in determining multi-receptor SARs on the basis of currently available compound data and can be further explored.

### Adenosine receptor antagonists

In Figure S1, the complete MTAL of the large adenosine receptor antagonist data set is shown. In this case, distinct clustering of patterns is also observed and the landscape representation identifies several centers of marked discontinuity. One of these centers is outlined in Figure S1 and shown in detail in Figure 4A. Here, systematic trends can also be identified. The potencies of all 10 compounds comprising this center of discontinuity are very similar against two of three receptors ('1_2'), but greatly differ against receptor 3. It follows that this region does not contain notable activity cliffs because the compounds display very similar potency patterns. However, in this case, the obvious discontinuity in activity space directly points at compound selectivity determinants. As shown in Figure 4B, these compounds also form an analog series and are only distinguished by R-groups at a single site of a phenyl moiety (in one instance, the phenyl ring is replaced with a pyridine). These substitutions differentiate one adenosine receptor subtype from two others, leading to potency changes of two to more than four orders of magnitude.

### Serine protease inhibitors

In Figure S2, the complete MTAL of the inhibitor set with activity against five related serine proteases is displayed. In this smaller data set, many compounds have similar (green) potency patterns. Two regions are outlined, region 1 with little apparent discontinuity and region 2 the most discontinuous region in this set. Region 1 is formed by four compounds mapping to a single neuron and shown in Figure 5A. Region 2 includes three neighboring neurons with a total of nine compounds (Figure 5B). These two regions contain two similar yet distinct series of analogs, shown in Figure 5C, D, respectively. The analogs in Figure 5C are distinguished by bioisosteric replacements at a single site, consistent with the presence of only low to moderate discontinuity in the activity landscape. By contrast, the analog series in Figure 5D is characterized by substitutions at two sites including bioisosteric replacements in para-position at the piperidine ring and less conservative substitutions at the sulfonamide group. In this case, large changes in potency profiles are observed, leading to the formation of activity cliffs and the introduction of substantial discontinuity in multi-target activity space (Figure 5B).

**Figure 3:** Multi-target activity landscape of the 4OR data set. (A) The complete activity landscape representation of 4OR is shown. Here, each compound is active against four targets and hence represented as an array of six target pair potency differences. A region with neurons representing a high level of multi-target structure-activity relationships discontinuity is outlined in red. (B) The three adjacent neurons outlined in (A) are shown. (C) The nine compounds assigned to these three neurons are shown, which represent a series of analogs. Structural modifications at specific substitution sites are highlighted in gray. Compounds are labeled with their source neurons referring to the respective neuron position in the grid representation of the complete activity landscape (e.g., [1 0 1]).

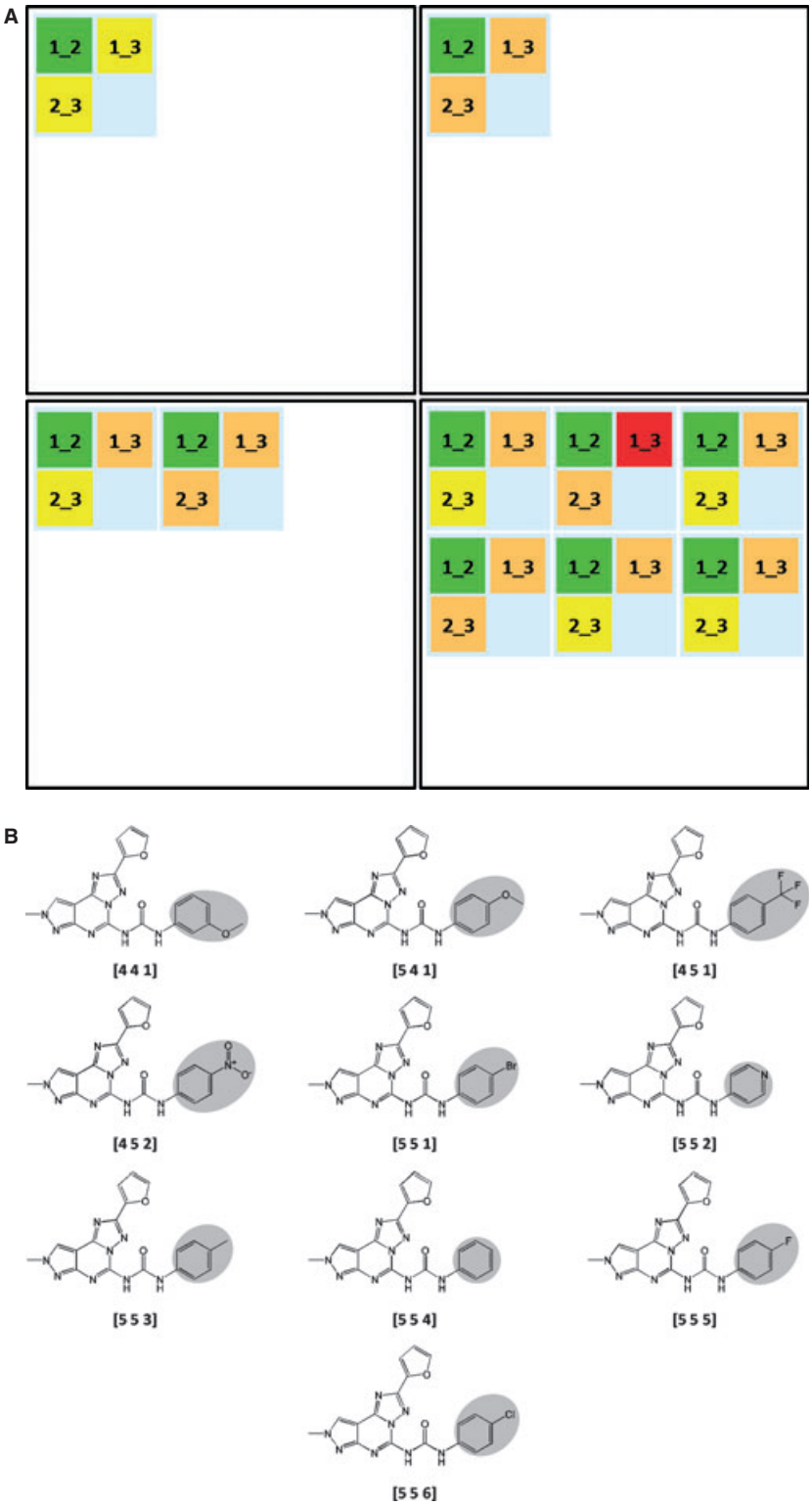



**Figure 4:** Subset of the activity landscape of 3AR. (A) Neurons forming a region of high discontinuity are shown that is outlined in red in Figure S1. (B) Compounds forming this region are shown. The representation is according to Figure 3C.

## Conclusions

Herein, we have reported the design of a SOM-based MTAL model that is, different from other activity landscape representations, primarily focused on the identification of discontinuity in multi-target activity space. For this purpose, the target pair potency-based compound representation introduced herein is a key design element. From compound subsets forming regions of high discontinuity, substitution sites and R-group patterns can be deduced that substantially influence multi-target SARs. Different compound data sets with three to five target annotations have been analyzed to illustrate the design and analysis principles. The concept of discontinuity in multi-target activity that is

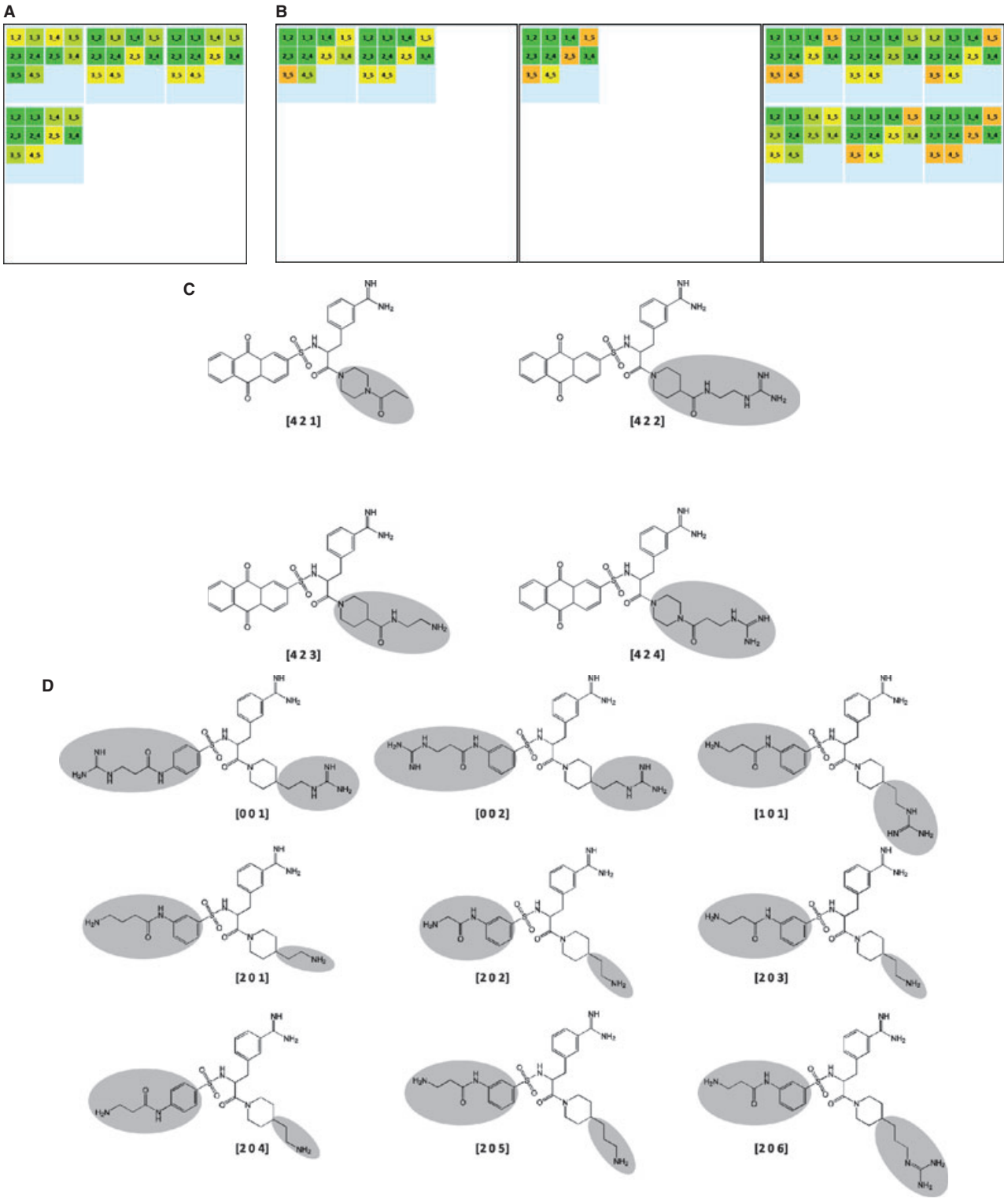



**Figure 5:** Subsets of the activity landscape of 5PR. (A) Neurons forming a region of low discontinuity (region 1 in Figure S2) are shown. (B) Another highly discontinuous region is shown (region 2 in Figure S2). (C) Compounds forming region 1 in (A) are shown. (D) Compounds forming region 2 in (B) are shown. The representation is according to Figure 3C.

embedded in the MTAL design is distinct from, yet complementary to, the study of SAR discontinuity in activity landscapes, which is associated with the presence of activity cliffs. As we have demonstrated, the presence of discontinuity in activity space might or might not correlate with the presence of activity cliffs. Among other aspects, the analysis of discontinuity in multi-target activity space is often critical for rationalizing compound selectivity. The newly introduced MTAL model is only the second representation of multi-target landscapes reported thus far and conceptually unique. As such, the MTAL design reported herein further extends the spectrum of available activity landscape representations and the knowledge that can be derived from them.

## Acknowledgments

The authors are grateful to Johann Gasteiger and Molecular Networks GmbH for providing the SONNIA software.

## 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. Maggiora G., Lajiness M., Bajorath J., Organizers. 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.
3. Wassermann A.M., Wawer M., Bajorath J. (2010) Activity landscape representations for structure-activity relationship analysis. J Med Chem;53:8209–8223.
4. 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.
5. Guha R., Van Drie J.H. (2008) Structure-activity landscape index: identifying and quantifying activity cliffs. J Chem Inf Model;48:646–658.
6. Wawer M., Peltason L., Weskamp N., Teckentrup A., Bajorath J. (2008) Structure-activity relationship anatomy by network-like similarity graphs and local structure-activity relationship indices. J Med Chem;51:6075–6084.
7. Shanmugasundaram V., Maggiora G.M. (2001) Characterizing property and activity landscapes using an information-theoretic approach. Proceedings of 222nd American Chemical Society National Meeting, Division of Chemical Information, Chicago, IL, August 26–30, 2001; Washington, D.C.: American Chemical Society; abstract no. 77.
8. Perez-Villanueva J., Santos R., Hernandez-Campos A., Giulianotti M.A., Castillo R., Medina-Franco J.L. (2011) Structure-activity relationships of benzimidazole derivatives as antiparasitic agents: Dual Activity-Difference (DAD) Maps. Med Chem Comm;2:44–49.
9. Yongye A.B., Byler K., Santos R., Martinez-Mayorga K., Maggiora G.M., Medina-Franco J.L. (2011) Consensus models of activity landscapes with multiple chemical, conformer, and property representations. J Chem Inf Model;51:1259–1270.
10. 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.
11. Hopkins A.L. (2008) Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol;4:682–690.
12. Keiser M.J., Irwin J.J., Shoichet B.K. (2010) The chemical basis of pharmacology. Biochemistry;49:10267–10276.
13. Mestres J., Gregori-Puigjané E. (2009) Conciliating binding efficiency and polypharmacology. Trends Pharmacol Sci;30:470–474.
14. Bajorath J. (2008) Computational analysis of ligand relationships within target families. Curr Opin Chem Biol;12:352–358.
15. Wassermann A.M., Peltason L., Bajorath J. (2010) Computational analysis of multi-target structure-activity relationships to derive preference orders for chemical modifications toward target selectivity. ChemMedChem;5:847–858.
16. Peltason L., Hu Y., Bajorath J. (2009) From structure-activity to structure-selectivity relationships: quantitative assessment, selectivity cliffs, and key compounds. ChemMedChem;4:1864–1873.
17. 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.
18. 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.
19. Rogers D., Hahn M. (2010) Extended-connectivity fingerprints. J Chem Inf Model;50:742–754.
20. Zupan J., Gasteiger J. (1999) Neural Networks in Chemistry and Drug Design, 2nd edn. Weinheim: Wiley-VCH.
21. Wassermann A.M., Dimova D., Bajorath J. (2011) Comprehensive analysis of single- and multi-target activity cliffs formed by currently available bioactive compounds. Chem Biol Drug Des;78:224–228.

## Notes

[a] *ChEMBL*; European Bioinformatics Institute (EBI): Cambridge, 2011, http://www.ebi.ac.uk/chembl/.
[b] Scitegic Pipeline Pilot; Accelrys Inc., San Diego, CA, USA, http://accelrys.com/products/scitegic/index.html.
[c] SONNIA – Self-Organizing Neural Network Package, Molecular Newtorks GmbH – Computerchemie, Erlangen, Germany, 2011, http://www.molecular-networks.com/products/sonnia.

## Supporting Information

Additional Supporting Information may be found in the online version of this article:

**Figure S1.** Multi-target activity landscape of the 3AR data set.

**Figure S2.** Multi-target activity landscape of the 5PR data set.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

## Summary

A new multi-target activity landscape representation has been reported in this chapter. Pair-wise similarities obtained from molecular fingerprints for data set compounds were projected into 2D space with the help of SOMs. Thus, the compounds were organized into a grid-like landscape. Individual compounds were displayed as arrays of pair-wise target activity relationships over all the targets. Such an arrayed arrangement of compounds was suitable for the identification of multi-target discontinuous regions. Inspection of ligand subsets that formed these centers of discontinuity provides important insights into substructural changes responsible for differential activity in multi-target activity space. It should also be noted that multi-target discontinuity is different from SAR discontinuity in single-target activity landscapes.

The landscape model described herein provides a unique perspective for studying SAR in multi-target activity space. Nevertheless, such models cannot be generated for ligands with incomplete activity annotations. In addition, interpretation of these landscapes with increasing numbers of targets becomes difficult. These limitations have been addressed in a second multi-target landscape representation reported in the following chapter.

# Chapter 7

# Navigating high-dimensional activity landscapes: design and application of ligand-target differentiation map

## Introduction

Compound profiling experiments serve as fundamental sources of potency information in the field of chemical biology and medicinal chemistry where compound libraries are simultaneously tested against several biological targets [1]. G protein coupled receptors and protein kinases are considered important drug targets due to their involvement in many physiological and biochemical processes [2, 3]. Therefore, profiling approaches often involve members of these therapeutically relevant protein families [4, 5]. Compound sets evaluated during profiling analyses can either show structural diversity or be chemically homogeneous. Moreover, the panels of selected targets often constitute representative members of the chosen target family. The outcome of these studies consists of a data matrix that contains multiple activity annotations for individual compounds.

Rationalization of these multi-dimensional bioactivity spaces and their exploration can be challenging. Nevertheless, such activity spaces provide a wide

variety of information regarding the compound activity profiles and ligand-target interaction patterns. The knowledge obtained from these analyses can often be utilized for the identification of molecular probes that differentiate between related members of a target family [6] as well as in compound design and optimization processes [5].

The activity landscape concept is often utilized for generating SAR models that combine pairwise similarity and potency difference distribution within compound data sets [7]. These landscape representations aid in SAR analysis as they are interpretable and intuitive in nature. However, these landscapes primarily focus on compound sets with activity against single targets. Molecular network-based activity landscape models for two [8] or more targets [9] have been recently introduced to assess the degree of selectivity within active ligands or the ability to form multi-target activity cliffs.

In addition to the originally reported network-based multi-target activity landscape representation, chemical similarity and activity difference maps generated using compound activity versus structural similarity plots [10] and models based on SOMs described in the previous section [11] have recently been reported. Nevertheless, these multi-target landscape representations account for activity data derived from limited number of targets (i.e. up to five) in an interpretable manner. Formalisms based on the activity landscape concept that describe actual high-dimensional activity spaces involving over 50 or more targets have yet to be introduced.

Availability of compound profiling data in the public domain is a limitation for the design of such multi-dimensional activity landscapes. Large scale profiling experiments are usually carried out by pharmaceutical companies for many therapeutically implicated targets, but the results are considered proprietary. However, a profiling study from Abbott Laboratories has recently been made public [12]. With the help of the resulting data set, design of a high-dimensional activity landscape representation has been attempted. This study has been published in reference [13] of this chapter. My contributions to the study reported herein have been to aid in the design of the novel multi-dimensional activity landscape representation and its implementation [13].

# Methodology

## Data Set

Metz et al. analyzed 3858 compounds against 172 diverse kinases spanning the kinome with the objectives of generating a kinase interaction map and analyzing various polypharmacological patterns [12]. A part of these results containing the chemical structures and bioactivity information for 1496 compounds was reported publicly. From this inhibitor set, 1473 compounds with unique 2D molecular representations [14] were retained.

Activity information for these 1473 compounds against all the 172 kinases was available in the form of negative logarithms of inhibitory constant endpoints or thresholds values. The activity matrix for these compounds was incomplete, i.e. of 172 kinases, pKi values for one to 122 targets were available for these compounds. The maximum overlap between the activity profiles of individual compounds was detected for 101 kinases. The incomplete yet multi-dimensional nature and several instances of partial overlaps among compound activity profiles were the challenging aspects that needed to be addressed during the activity landscape design. In this case, a difference of one order of magnitude between pKi values was set as the activity threshold.

## Assessment of Chemical Similarity

Incorporation of pairwise structural similarity among the constituent compounds within a ligand data set is a fundamental requirement for the generation of activity landscape representations. During the analysis, two conceptually different approaches were utilized to evaluate chemical similarities between the data set compounds. MACCS structural keys [15], a molecular fingerprint representation was used to calculate pairwise whole molecule similarities using the Tanimoto coefficient (Tc) [16] as the similarity measure. Compound pairs with a Tc value of at least 0.8 were considered to be structurally similar. Additionally, MMP analysis was applied to assess direct substructural changes within compounds [17]. Thus, all compound pairs forming MMPs, i.e. differing by single site chemical modifications, were identified.

## Motivation for Novel Design

Activity landscape representations combine the information originating from systematic pairwise structural similarities and potency differences within ligand sets. Thus, in addition to assessing single and multi-target SAR relevant features, a high-dimensional activity landscape model must provide complete coverage of activity data as well as handle its sparseness.

Initial attempts were made to assess activity similarity using nonbinary Tc (nbTc) [18] for all compound pairs. The nbTc value was calculated using the activity data for all targets shared by a given compound pair $(A, B)$ as:

$$nbTc(A,B) = \frac{\sum_{i=0}^{n} pot_A^i * pot_B^i}{\sum_{i=0}^{n} [(pot_A^i)^2 + (pot_B^i)^2 - pot_A^i * pot_B^i]}$$

where $nbTc$ denotes the activity similarity value while $pot_A^i$ and $pot_B^i$ represent the potencies of the $i^{th}$ shared target for compounds $A$, $B$, respectively. Systematic pairwise comparison of activity similarities within the data set revealed that a single nbTc value corresponded to a range of overlapping targets while a given number of shared targets was associated with several nbTc values. Examination of nbTc distribution for shared targets with more than one order of magnitude potency difference showed that high similarity values were obtained by small number of shared targets exceeding the ten fold activity difference and low activity similarity was observed when the number of shared targets with significant activity differences was high. These observations were consistent with the expectation associated with a similarity measure. However, several instances were also observed when high nbTc values were obtained from large number of shared targets with potency differences greater than one order of magnitude and low nbTc values were produced by small number of overlapping targets with more than ten fold activity difference. The high-dimensional activity landscape design also needed to address this ambiguity observed with respect to the calculated activity similarities.

## Ligand-Target Differentiation Map

A novel activity landscape representation, the ligand-target differentiation (LTD) map, has been designed keeping in mind the aforementioned objectives. The core elements of LTD map are compound pairs and these are analyzed from three perspectives [13]. Firstly, numbers of compound pairs with varying degrees of target overlaps are reported. As a second aspect, the numbers of compound pairs within the first subset having more than ten fold activity differences for the overlapping targets are analyzed. Structurally similar compound pairs within the first two subsets comprise the third layer of information. The LTD map and its various design elements are reported in [13].

A grid representation forms the framework of the LTD map. Bivariate data containing the numbers of overlapping targets and those with potency difference over one order of magnitude are plotted along x- and y- axes, respectively. These values are binned to produce unit cells with a constant dimension of five by five. Therefore, these squared unit cells account for the entire range of possible pairwise target and target activity difference relationships underlying the data set. Compound pair frequencies present within the cell bounds are calculated and a color spectrum from light pink over magenta to black is applied to assess the magnitude of their occurrence. Thus, black colored cells contain one compound pair while those with increasing numbers are shown in various shades of magenta and the cell containing the largest number of pairs is depicted in lightest pink color. Absence of any compound pairs within respective data intervals produce "empty" cells [13].

Structural similarity information has been added using inlays. The frequency of compound pairs participating in structural relationships is monitored by an inverse color spectrum from light(est) blue over dark blue to black. Thus, lighter shades of blue correspond to smaller number of chemically related compound pairs and darker blue shades indicate larger frequencies while the black cell contains the highest number of compound pairs. For the Abbott data set, LTD maps were generated for two complementary similarity assessment approaches, MACCS Tc and MMP, respectively, and shown in **Figure 1a** and **1b**. The LTD maps were implemented in R environment [19].

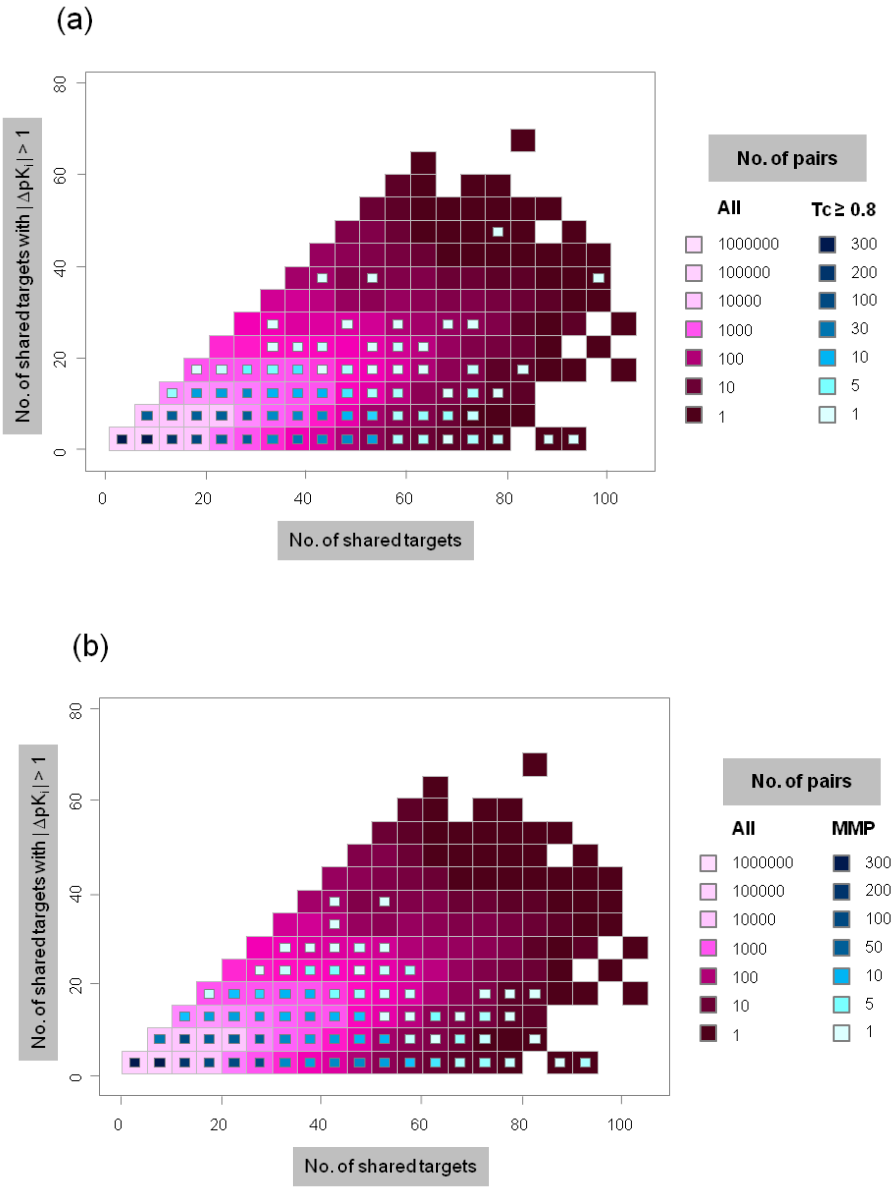



**Figure 1: LTD map for kinase inhibitor data set.** Two LTD maps for the data set consisting of 1473 kinase inhibitors using alternate structural similarity assessment approaches are shown. Whole molecular similarity relationships determined using MACCS Tc are illustrated in (a) while substructure based relationships obtained using MMP analysis are depicted in (b). Compound pairs participating in MMP formation share a common core structure. *(taken from Iyer et. al[13])*

# Results

As shown in **Figure 1**, information associated with high-dimensional activity space can be readily viewed using the LTD map [13]. For the kinase inhibitor

data set with activity annotations against 172 kinases, the maximum target overlap between individual compound pairs was found to be 101 kinases and for 69 kinases, the potency difference was greater than one order of magnitude. Majority of the compound pairs were found to lie within the data intervals of 0 to ∼20 shared targets and 0 to ∼10 shared targets with qualifying activity differences, as portrayed by cell colors. Furthermore, it was also observed that the number of compound pairs showed a sharp decline with rise in the interval of shared targets and shared targets with significant potency differences. Cells comprising the leftmost (pseudo-diagonal) region consisted of compound pairs for which almost all the targets showed significant differences in activity while cells spanning the bottom rows displayed small activity variations. Additional information regarding the pairwise similarity relationships with respect to MACCS Tc and substructure equivalences determined by MMP, was obtained by the examination of the inlays in **Figure 1a** and **1b**, respectively. It was noted that such similarity relationships were prevalent within compounds having low target overlap. Moreover, chemically similar compound pairs with up to ∼50 shared targets and ∼10 shared targets with potency differences above the activity threshold were also relatively frequent. Closely related compounds with similar potency against all kinase targets were found in the cells forming lower rows of the LTD map. Different from these pairs, compound pairs that were structurally similar but showed significant deviation in activity against nearly all the shared targets were associated with the diagonally situated cells. The cells located in the upper right rows consisted of structurally unrelated compound pairs as inlays were rarely observed. Thus, structural similarity and activity distributions in multi-dimensional activity space could be navigated with the help of the LTD map [13].

## Summary

The ligand-target differentiation map, a novel data structure has been introduced that integrates high-dimensional activity data with chemical similarity relationships present between various compounds in a systematic manner. Thus, the LTD map constitutes a multi-dimensional graphical activity landscape model providing access to multi-target SAR information content. Limited public availability of profiling results has been an important challenge faced during the design of high-dimensional landscape representations. Recently, a kinase inhibitor data set containing over 1400 compounds evaluated against 172 kinases was made publicly accessible. Subsequent analysis made possible by this data set led to the design of the LTD map. The scaling down of the complexity associated with incomplete multi-dimensional activity information serves as the key element of the computational methodology described in this analysis. This simplification has been achieved by systematically determining pairwise differences between overlapping targets. Further, bivariate binning of the numbers of shared targets and shared targets with significant activity differences to form unit cells of constant size has been carried out for the purpose of visualization. Addition of pairwise chemical similarities completes the landscape view that facilitated navigation of various multi-target SAR relevant regions. The relative ease of generation and flexibility in application to lower dimensional activity spaces make LTD map an invaluable tool to carry out multi-target SAR analysis.

The study reported herein has been published in reference [13] of this chapter. My contributions to this study have been to aid in the design of the LTD map and its implementation [13].

## References

[1] Rix U., Superti-Furga G. Target profiling of small molecules by chemical proteomics. *Nature Chem. Biol.*, **2009**, 5, 616-624.

[2] Arinaminpathy Y., Khurana E., Engelman D. M., Gerstein M. B. Computational analysis of membrane proteins: the largest class of drug targets. *Drug Discov. Today*, **2009**, 14, 1130-1135.

[3] Cohen P. Protein kinases — the major drug targets of the twenty-first century? *Nat. Rev. Drug. Discov.*, **2002**, 1, 309-315.

[4] 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.

[5] 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.

[6] Bajorath J. Computational analysis of ligand relationships within target families. *Curr. Opin. Chem. Biol.*, **2008**, 12, 352-358.

[7] Wassermann A. M., Wawer M., Bajorath J. Activity landscape representations for structure-activity relationship analysis. *J. Med. Chem.*, **2010**, 53, 8209-8223.

[8] Peltason L., Hu Y., Bajorath J. From structure-activity to structure-selectivity relationships: Quantitative assessment, selectivity cliffs, and key compounds. *Chem. Med. Chem.*, **2009**, 4, 1864-1873.

[9] 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.

[10] Medina-Franco J. L., Yongye A. B., Perez-Villanueva J., Houghten R. A., Martínez-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] Iyer P., Dimova D., Vogt M., Bajorath J. Navigating high-dimensional activity landscapes: design and application of ligand-target differentiation map. *J. Chem. Inf. Model*, **2012**, 52, 1962-1969.

[14] 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.

[15] *MACCS Structural Keys*, Symyx Software: San Ramon, CA, USA.

[16] Willett P., Barnard J. M., Downs G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.*, **1998**, 38, 983-996.

[17] Wassermann A. M., Dimova, D., Iyer P. and Bajorath J. Advances in computational medicinal chemistry: matched molecular pair analysis. *Drug Develop. Res.*, **2012**, 73, 518-527.

[18] Comparison of nonbinary similarity coefficients for similarity searching, clustering and compound selection. *J. Chem. Inf. Model.*, **2009**, 49, 1193-1201.

[19] *R: A Language and Environment for Statistical Computing*, R Development Core Team, R Foundation for Statistical Computing: Vienna, Austria, **2008**.

In the following chapter, analysis performed using the data obtained from a second profiling experiment has been reported.

# Chapter 8

# Assessing the target differentiation potential of imidazole-based protein kinase inhibitors

## Introduction

Compound profiling experiments to evaluate ligand sets against a panel of protein targets are frequently performed for comprehensive characterization of ligand-target interactions. Such studies are prominent sources of multi-target activity information and provide valuable insights during SAR and selectivity analyses. Systematic exploration of activity data obtained from a profiling experiment consisting of 484 imidazole-based inhibitors tested against 24 different kinases was carried out to identify compounds with high differentiation potential. The differentiation potential served as a measure to evaluate the ability of these inhibitors to distinguish between the various kinases.

Journal of Medicinal Chemistry

Brief Article

pubs.acs.org/jmc

# Assessing the Target Differentiation Potential of Imidazole-Based Protein Kinase Inhibitors

Dilyana Dimova,[†,∥] Preeti Iyer,[†,∥] Martin Vogt,[†] Frank Totzke,[§] Michael H. G. Kubbutat,[§] Christoph Schächtele,[§] Stefan Laufer,*[,‡] 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

[‡]Department of Pharmacy and Biochemistry, Pharmaceutical/Medicinal Chemistry, Eberhard-Karls-Universität Tübingen, Auf der Morgenstelle 8, D-72076 Tübingen, Germany

[§]ProQinase GmbH, Breisacher Strasse 117, D-79106 Freiburg, Germany

Supporting Information

**ABSTRACT:** A library of 484 imidazole-based candidate inhibitors was tested against 24 protein kinases. The resulting activity data have been systematically analyzed to search for compounds that effectively differentiate between kinases. Six imidazole derivatives with high kinase differentiation potential were identified. Nearest neighbor analysis revealed the presence of close analogues with varying differentiation potential. Small structural modifications of active compounds were found to shift their inhibitory profiles toward kinases with different functions.

## INTRODUCTION

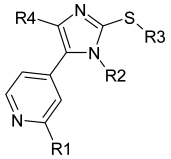
Profiling of compound collections against target families is an important source of activity data for chemical biology and drug discovery.[1] Profiling has become a popular approach to characterize ligand-based relationships between targets[2] and identify new active compounds, especially for high-profile therapeutic targets such as G protein coupled receptors[3,4] or protein kinases.[5,6] Target profiling experiments are frequently carried out in pharmaceutical research environments, but these proprietary results are rarely disclosed, with occasional exceptions.[7,8] Exemplary profiling studies have substantially advanced our understanding of structure−activity relationships (SARs) and selectivity patterns within important target families. For example, profiling of kinase inhibitors against different subfamilies of the kinome revealed unexpected cross-reactivity of many kinase inhibitors,[7] hence providing insights into polypharmacological behavior of clinically relevant inhibitors. In addition, molecular network analysis has been applied to analyze kinase profiling data and rationalize activity patterns.[8] On the basis of kinase profiling data,[8] matched molecular pair analysis has also been carried out to propose inhibitors with increased kinase selectivity.[9] However, kinase profiling is a laborious and expensive part of medicinal chemistry programs, as it requires large assay efforts and high costs. Consequently, in silico support or guidance in study design and data analysis, even if approximate, should be of considerable help for the community. Several computational studies have analyzed available kinase activity data. For example, machine learning models have been derived to search for kinase inhibitors on a large scale[10] or process profiling data and predict cross-reactivity of kinase inhibitors.[11]

Herein, we report a kinase profiling experiment using a library of imidazole-based adenosine triphosphate (ATP) site-directed kinase inhibitors. Different from previous investigations, much emphasis has been put on the exploration of kinase differentiation potential of candidate inhibitors. The concept of kinase differentiation potential is distinct from kinase selectivity of inhibitors. Compounds with differentiation potential must display significantly varying potency levels against multiple kinases.

## METHODS AND MATERIALS

**Kinases, Inhibitors, and Profiling Assays.** A set of 484 pyridinylimidazole based inhibitors with general structure **I** (Scheme 1) were tested for kinase inhibition using 24 different kinases (AKT1, ARK5, Aurora-A, Aurora-B, BRAF VE, CDK2/CycA, CDK4/CycD1, COT, AXL, EGF-R, EPHB4, ERBB2, FAK, IGF1-R, SRC, VEGF-R2, CK2-$\alpha$1, JNK3, MET, p38-$\alpha$, PDGFR-$\beta$, PLK1, SAK, TIE2). These kinases were selected because they are implicated in different forms of cancer. The 484 different derivates were synthesized and characterized (including their purity) as described previously.[12−17] Kinase activity data were generated with the ProQinase free choice biochemical kinase assay system. Activities were determined as residual activities (% of control).[18]

**Scheme 1**




General structure I

Initially, compounds were screened at a single concentration of 10 $\mu$M. Subsequently, titration curves were generated for clearly active compounds. Then the coefficient of variation (CV) between the initial

Received: October 6, 2012
Published: December 4, 2012

   dx.doi.org/10.1021/jm3014508 | J. Med. Chem. 2012, 55, 11067−11071

screen and subsequent assays was determined for each kinase. The average CV was only 7.7% (for only 3 of 24 kinases, values of 10−12% were obtained), thus indicating that activity data for the initial single-point experiments were reliable.

**Analysis of Kinase Differentiation Potential.** Residual activities for single-point measurements were logarithmically transformed into a numerically stable data format for subsequent analysis, as illustrated in Figure 1. According to this transformation, a logarithmic value of 2

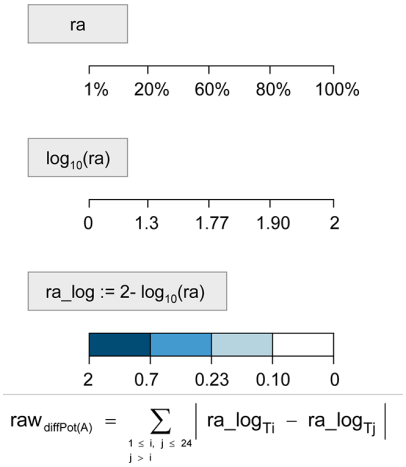



**Figure 1.** Scoring scheme. Measured residual activities (ra) were initially converted into logarithmic values used for the calculation of the raw differential potential ($raw_{diffPot}$) of each compound. Logarithmic values were adjusted such that 2 indicated (nearly) complete inhibition and 0 no inhibition and aligned with the original experimental binning scheme. Color code is as follows: dark blue, ≤20%; blue, >20%, ≤60%; light blue, >60%, ≤80%; white >80% residual wild-type activity.

indicates (nearly) full inhibition and a value of 0 no inhibition. On the basis of these transformed activity values, a raw target differentiation potential score was calculated as follows (see also Figure 1):

$$\mathrm{raw}_{\mathrm{diffPot(A)}} = \sum_{\substack{1\le i,j\le 24 \\ j>i}} |\mathrm{ra\ log}\ T_i - \mathrm{ra\ log}\ T_j|$$

Here, the logarithmic terms refer to the transformed activity of a compound to targets $T_i$ and $T_j$, respectively. For each compound, all possible target pairs were formed and activity differences were summed. Thus, according to this formalism, compounds have high differentiation potential if they display large activity differences against many target pairs. Raw scores were then transformed into standard $Z$-scores and normalized through mapping onto a cumulative distribution function assuming a normal distribution, yielding final scores between 0 (lowest differentiation potential) and 1 (highest potential). This scoring scheme represents a further refined and generalized version of a binned cumulative differentiation score previously used to characterize ligands of different target families.[19]

**Nearest Neighbor Analysis.** For selected active compounds, nearest structural neighbors were identified on the basis of systematic pairwise comparisons. For this purpose, Tanimoto similarity[20] was calculated using MACCS structural keys[21] as a molecular representation. As a nearest neighbor criterion, a threshold value of more than 80% MACCS Tanimoto similarity was applied.

**Activity Profiles.** For preferred inhibitors and their nearest neighbors, activity profiles were generated using the activity-based color code shown in Figure 1. In these profiles, each bin corresponds to the activity against a specific kinase.

## RESULTS AND DISCUSSION

**Kinase Inhibitor Data.** The complete matrix reporting activities for all 484 compounds against the 24 kinases is provided in Table S1 of the Supporting Information. All residual activities were transformed into a logarithmic format (as described above) and subjected to computational analysis.

**Compound Differentiation Potential.** Figure 2 shows the distribution of normalized $Z$-scores for all test compounds.

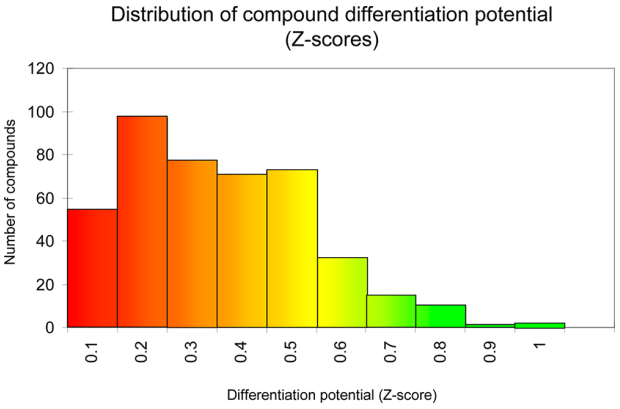



**Figure 2.** Distribution of compound differentiation potential. The histogram shows the distribution of $Z$-scores for all test compounds. $Z$-scores were normalized to the value range between 0 (lowest differentiation potential) and 1 (highest potential) and binned on the $X$-axis into 10 equally sized score intervals. The $Y$-axis reports the number of compounds falling into each interval. The differentiation potential of the compounds (normalized $Z$-scores) was color-coded using a spectrum ranging from red (lowest differentiation potential) over yellow (intermediate) to green (highest differentiation potential).

The score distribution directly reflects the kinase differentiation potential of the inhibitors. The distribution reveals that most of the compounds fell within the range of low (red) to intermediate (yellow) differentiation potential, as one might expect for ATP site-directed inhibitors. However, the distribution also contained a notable tail toward high (green) differentiation potential. Hence, a small subset of test compounds displayed a much higher than average potential to differentiate between the selected cancer-relevant kinases.

**Preferred Inhibitors.** On the basis of the score distribution in Figure 2, we selected the imidazole derivatives with the highest differentiation potential, falling into the scoring interval [0.78, 1.00]. The structures of these in part closely related analogues are shown in Figure 3 with their activity profiles.

In the next step, nearest structural neighbors of each of the six top-scoring compounds were identified in the data set and their differentiation potential was compared, as reported in Figure 4. Here, notable differences were observed. For example, the top-scoring inhibitor with the highest differentiation potential had only one nearest neighbor, the fourth-ranked compound (Figure 4a). Equivalent observations were made for inhibitors at rank 3 (Figure 4c) and 4 (Figure 4d). By contrast, the inhibitor at rank 2 (Figure 4b) had a total of 12 nearest structural neighbors with variable differentiation potentials. Similar observations were made for inhibitors at ranks 5 (Figure 4e) and 6 (Figure 4f), having four and six neighbors, respectively. These compounds also displayed low to intermediate differentiation potential. From these compound series, SAR patterns emerged, as discussed in the following.


dx.doi.org/10.1021/jm3014508 | J. Med. Chem. 2012, 55, 11067−11071

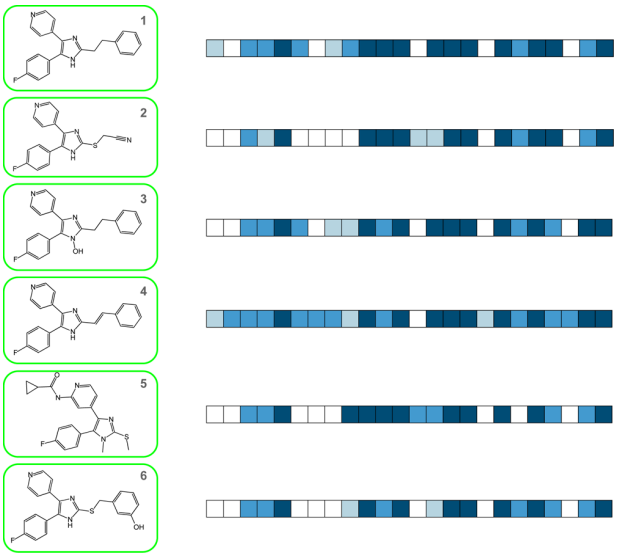


**Figure 3.** Top-ranked kinase inhibitors. Shown are the six top-ranked compounds with highest differentiation potential (labeled with their ranks) together with their activity profiles (color-coded according to Figure 1). In the activity profile, each bin is assigned to one of the 24 kinases.

**SAR Analysis.** All tested compounds were initially designed as potential p38$\alpha$ MAP kinase inhibitors. The major novelty of these imidazole-based series is the 2-thio substitution, which greatly reduces their ability to bind to cytochrome P450 (CYP) enzymes by complexing the iron in the active site. This CYP interaction presented a general problem associated with first-generation imidazole-based inhibitors. Kinase profiles of a large set of structurally closely related inhibitors have not yet been described. However, the results reported herein demonstrate how even minor structural modifications of closely related inhibitors can alter the inhibition profile toward kinases other than p38, including representatives of kinase families with rather different functions such as receptor tyrosine kinases.

The computational approach designed for the analysis of the kinase profiling matrix did not take structural information about the kinase ATP binding site into account. Nevertheless, it detected activity differences between compounds that were consistent with structural data of p38−inhibitor interactions. Figure 5 shows an outline of p38 bound to the ATP site-directed pyridinylimidazole inhibitor SB203580,[22] as revealed by the X-ray structure of the complex.[23] A critically important hydrogen bond is formed between the pyridin-4-yl group and the backbone NH of Met109. Another hydrogen bond is formed between Lys53 and N-3 of the imidazole core. In addition, there is a $\pi-\pi$ stacking between Tyr35 and the phenyl ring of the inhibitor. The 4-fluorophenyl ring is accommodated in hydrophobic region I, while hydrophobic region II is not occupied. On the basis of these interaction patterns, structural modifications of imidazole-based inhibitors that led to changes in their differentiation potential according to Figure 4 can be rationalized. For example, the compounds in Figure 4b,c very well reflect the relevance of the $\pi-\pi$ interaction of the S-residues at the R3 position with Tyr35 in p38, as indicated in Figure 5. In Figure 4b, **2** with an acetonitrile group at this position had overall highest differentiation potential, whereas smaller or larger (aromatic) substituents at this position led to a gradual loss of this potential. Phenyl-based substituents such as

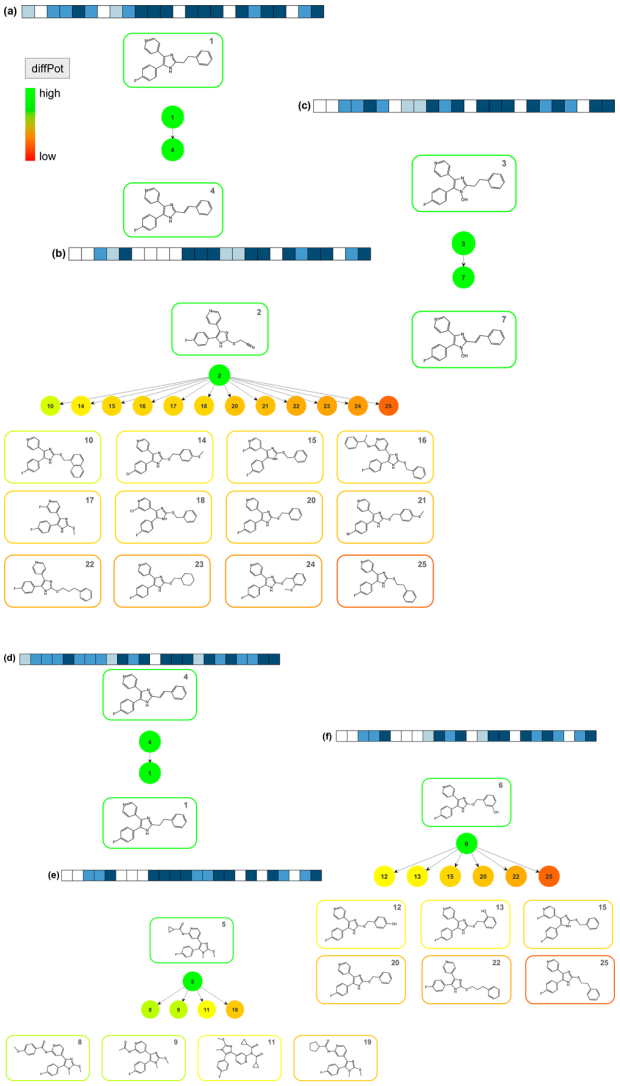



**Figure 4.** Top-ranked inhibitors and nearest neighbors. In (a) to (f), the six compounds with highest differentiation potential are shown together with their nearest structural neighbors (i.e., all other compounds having at least 80% 2D structural similarity). In the center, each of the six top-ranked compounds is represented as the root node and nearest neighbors form leaves. The nodes are color-coded according to differentiation potential as in Figure 2. The activity profile of the root compound is displayed, and compound structures are drawn proximal to their nodes. Multiple nearest neighbors are arranged according to decreasing differentiation potential from the left to the right.

$\pi-\pi$ interactors were generally more difficult to accommodate than the acetonitrile group because they required a coplanar orientation for best interactions. As illustrated in Figure 4c, loss of R-group flexibility to adopt a favorable geometry for the $\pi-\pi$ interaction also resulted in a penalty and altered the differentiation potential observed for a compound with a conformationally unrestricted phenyl group.

In our kinase panel, JNK3 was most closely related to p38. The only difference in the ATP-binding site of these kinases is the gatekeeper residue, which is Thr in p38 and Met in JNK3. Met is larger but has a flexible side chain and can conformationally adapt. Given the similarity of these kinases,


dx.doi.org/10.1021/jm3014508 | J. Med. Chem. 2012, 55, 11067−11071

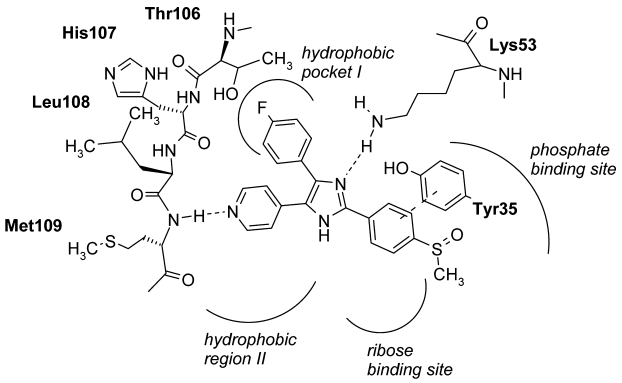



**Figure 5.** p38–inhibitor complex. Shown is the structure of the ATP binding site in p38 in complex with the pyridinylimidazole inhibitor SB203580.

many compounds inhibited them comparably. However, nearest neighbor analysis also revealed interesting exceptions. For example, **2** with its acetonitrile substituent was highly active against p38 and JNK3. By contrast, **15**, a structural neighbor of **2** with a phenyl group at the corresponding position, retained high activity against p38 but was not active against JNK3. Similarly, **25**, another structural neighbor with an additional methylene group in the linker presenting the phenyl substituent, showed reduced activity against P38 and was also inactive against JNK3. Both of these compounds had overall only low differentiation potential. In the panel, AKT1 was the kinase most distantly related to p38. Accordingly, many of the p38-directed compounds did not inhibit AKT1. However, there were exceptions among compounds with high differentiation potential. For example, **1** and **4** strongly inhibited p38 but also displayed weak activity against AKT1.

Furthermore, very small structural changes between compounds with high differentiation potential preferentially affected certain subsets of kinases. For example, **1** and **4** were only distinguished by the presence of a double bond in the linker between the imidazole core and a phenyl substituent (thus slightly reducing the conformational flexibility of **4**). This minute change led to overall higher activity of **4** against the kinase panel than **1**. In particular, it affected binding to cyclin-dependent kinases, against which **4** was active but **1** only weakly active or inactive. In addition, the presence of a hydrophilic group in this region of the inhibitors, for example, in **6**, led to a complete loss of activity against these kinases. Another interesting example was inhibition of PLK1. Among compounds with significant differentiation potential, only **4**, **7**, and **10** inhibited this kinase; all others were inactive. Compounds **4** and **7** were structurally highly similar, but in **10**, the conformationally restricted phenyl substituent was replaced by an unrestricted naphthalene group. Despite this change, the activity profiles of all three compounds were overall similar and distinct from many others.

**Differentiation Potential versus Selectivity.** Differentiation potential as assessed herein is related to but distinct from compound selectivity, for which other measures have been introduced in the kinase inhibitor field. These include, among others, the Ambit selectivity score[24] and the thermodynamic partition index.[25] The latter coefficient reflects the partitioning of inhibitor binding across a panel of kinases at thermodynamic equilibrium and should thus be calculated on the basis of equilibrium constants (i.e., $K_i$ or $K_d$). Hence, it is not applicable to residual activities or other approximate measurements. The Ambit score (AS) is calculated as the fraction of $n$ tested kinases that are inhibited by a compound at a given threshold value of residual activity. Hence, a score of 0 indicates a compound that is inactive at the selected threshold and a score of 1 a compound that is consistently active and nonselective. By contrast, a target-selective compound obtains a score of $1/n$ (close to 0). We have calculated AS values for all compounds for a threshold value of less than 60% residual activity, as reported in Table S2 of the Supporting Information. The mean and standard deviation of the AS distribution are 0.31 and 0.22, respectively. For **1**–**6** with the highest target differentiation potential, scores range from 0.54 and 0.83. Hence, these compounds would not be considered on the basis of simple selectivity scoring. At lower levels of residual activity (e.g., 30%), the scores consistently decrease and equivalent conclusions are drawn. These results reflect the conceptual difference between target differentiation potential and target selectivity of inhibitors. Compounds with differentiation potential are often rich in multitarget SAR information.

## CONCLUSIONS

Herein we have reported a compound profiling experiment on a set of cancer-relevant kinases using ATP site-directed imidazole derivatives, combined with a computational study to identify compounds with kinase differentiation potential. Several structurally closely related inhibitors with high differentiation potential were identified, and SAR features were explored on the basis of nearest neighbor analysis. In a number of instances, small structural modifications of closely related compounds led to substantial alterations of their inhibitory profiles, in part involving kinases with different functions. On the basis of these results, the evaluated compound series should merit further consideration in the development of selective kinase inhibitors. Furthermore, the computational approach reported herein is readily applicable to the analysis of other compound profiling experiments and the identification of active small molecules with target differentiation potential.

## ASSOCIATED CONTENT

### Supporting Information

Table S1 listing complete 484 × 24 kinase profiling matrix and Table S2 listing the distribution of AS values for all compounds. This material is available free of charge via the Internet at http://pubs.acs.org.

## AUTHOR INFORMATION

### Corresponding Author

*For S.L.: phone, +49-7071-2978788; fax, +49-7071295037; e-mail, Stefan.laufer@uni-tuebingen.de. For J.B.L phone, +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.

## ACKNOWLEDGMENTS

The authors are grateful to Dagmar Stumpfe for help with data preparation.

## ■ ABBREVIATIONS USED

AS, Ambit score; ATP, adenosine triphosphate; CV, coefficient of variation; CYP, cytochrome P450; diffPot, differentiation potential; SAR, structure−activity relationship

## ■ REFERENCES


(1) Rix, U.; Superti-Furga, G. Target Profiling of Small Molecules by Chemical Proteomics. *Nat. Chem. Biol.* **2009**, *5*, 616−624.

(2) Bajorath, J. Computational Analysis of Ligand Relationships within Target Families. *Curr. Opin. Chem. Biol.* **2008**, *12*, 352−358.

(3) Kenakin, T. Functional Selectivity in GPCR Modulator Screening. *Comb. Chem. High Throughput Screening* **2008**, *11*, 337−343.

(4) 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.

(5) Goldstein, D. M.; Gray, N. S.; Zarrinkar, P. P. High-Throughput Kinase Profiling as a Platform for Drug Discovery. *Nat. Rev. Drug Discovery* **2008**, *6*, 391−397.

(6) Bi, K.; Lebakken, C. S.; Vogel, K. W. Transformation of in Vitro Tools for Kinase Profiling: Keeping an Eye over the Off-Target Liabilities. *Expert Opin. Drug Discovery* **2011**, *6*, 701−712.

(7) Fabian, M. A.; Biggs, W. H., 3rd; Treiber, D. K.; Atteridge, C. E.; Azimioara, M. D.; Benedetti, M. G.; Carter, T. A.; Ciceri, P.; Edeen, P. T.; Floyd, M.; Ford, J. M.; Galvin, M.; Gerlach, J. L.; Grotzfeld, R. M.; Herrgard, S.; Insko, D. E.; Insko, M. A.; Lai, A. G.; Lélias, J. M.; Mehta, S. A.; Milanov, Z. V.; Velasco, A. M.; Wodicka, L. M.; Patel, H. K.; Zarrinkar, P. P.; Lockhart, D. J. A Small Molecule−Kinase Interaction Map for Clinical Kinase Inhibitors. *Nat. Biotechnol.* **2005**, *23*, 329−336.

(8) 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.

(9) Milletti, F.; Hermann, J. C. Targeted Kinase Selectivity from Kinase Profiling Data. *ACS Med. Chem. Lett.* **2012**, *3*, 383−386.

(10) Martin, E.; Mukherjee, P. Kinase-Kernel Models: Accurate in Silico Screening of 4 Million Compounds across the Entire Human Kinome. *J. Chem. Inf. Model.* **2012**, *52*, 156−170.

(11) Niijima, S.; Shiraishi, A.; Okuno, Y. Dissecting Kinase Profile Data To Predict Activity and Understand Cross-Reactivity of Kinase Inhibitors. *J. Chem. Inf. Model.* **2012**, *52*, 901−912.

(12) Laufer, S. A.; Wagner, G. K. From Imidazoles to Pyrimidines: New Inhibitors of Cytokine Release. *J. Med. Chem.* **2002**, *45*, 2733−2740.

(13) Laufer, S. A.; Kotschenreuther, D.; Wagner, G. K. Ones, Thiones and *N*-Oxides: An Exercise in Imidazole Chemistry. *Angew. Chem., Int. Ed.* **2002**, *41*, 2290−2293.

(14) Laufer, S. A.; Striegel, H.-G.; Wagner, G. K. Imidazole Inhibitors of Cytokine Release: Probing Substituents in the 2 Position. *J. Med. Chem.* **2002**, *45*, 4695−4705.

(15) Laufer, S. A.; Kotschenreuther, D. A.; Wagner, G. K.; Albrecht, W. Novel Substituted Pyridinyl Imidazoles as Potent Anti-Cytokine Agents with Low Cytochrome P450 Activity. *J. Med. Chem.* **2003**, *46*, 3230−3244.

(16) Laufer, S. A.; Kotschenreuther, D. A.; Wagner, G. K. Identification of Regioisomers in a Series of N-Substituted Pyridin-4-yl-imidazole Derivatives by Regiospecific Synthesis, GC/MS and $^{1}$H-NMR. *J. Org. Chem.* **2003**, *68*, 4527−4530.

(17) Laufer, S. A.; Striegel, H.-G.; Zimmermann, W.; Ruff, K. J. Tetrasubstituted Imidazole Inhibitors of Cytokine Release: Probing Substituents in the N-1 Position. *J. Med. Chem.* **2004**, *47*, 6311−6325.

(18) ProQinase Free Choice Biochemical Kinase Assays. http://www.proqinase.com/ (accessed September 3, 2012).

(19) Dimova, D.; Bajorath, J. Computational Chemical Biology: Identification of Small Molecular Probes that Discriminate between Members of Target Protein Families. *Chem. Biol. Drug Des.* **2012**, *79*, 369−375.

(20) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(21) *MACCS Structural Keys*; Symyx Software: San Ramon, CA, U.S., 2005.

(22) Wagner, G.; Laufer, S. Small Molecular Anti-Cykotine Agents. *Med. Res. Rev.* **2006**, *26*, 1−62.

(23) Tong, L.; Pav, S.; White, D. M.; Rogers, S.; Crane, K. M.; Cywin, C. L.; Brown, M. L.; Pargellis, C. A. A Highly Specific Inhibitor of Human p38 MAP Kinase Binds in the ATP Pocket. *Nat. Struct. Biol.* **1997**, *4*, 311−316.

(24) Karaman, M. W.; Herrgard, S.; Treiber, D. K.; Gallant, P.; Atteridge, C. E.; Campbell, B. T.; Chan, K. W.; Ciceri, P.; Davis, M. I.; Edeen, P. T.; Faraoni, R.; Floyd, M.; Hunt, J. P.; Lockhart, D. J.; Milanov, Z. V.; Morrison, M. J.; Pallares, G.; Patel, H. K.; Pritchard, S.; Wodicka, L. M.; Zarrinkar, P. P. A Quantitative Analysis of Kinase Inhibitor Selectivity. *Nat. Biotechnol.* **2008**, *26*, 127−132.

(25) Cheng, A. C.; John Eksterowicz, J.; Geuns-Meyer, S.; Sun, Y. Analysis of Kinase Inhibitor Selectivity Using a Thermodynamics-Based Partition Index. *J. Med. Chem.* **2010**, *53*, 4502−4510.

# Summary

Profiling data obtained after testing 484 ATP-site directed imidazole derivatives against 24 cancer-associated kinases was systematically analyzed to identify ligands with extremely variable inhibition profiles. Alterations in inhibition profiles were measured by quantifying activity differences over all possible target pairs expressed as compound differential potential. Six ligands with highest variation in their inhibitory profiles were identified after ranking compounds according to their differential potentials. Nearest neighbor analysis of these compounds revealed structurally related analogs with differential potentials of varying magnitude. Many instances where minor chemical modifications produced large variations in compound inhibitory profiles were also identified. Therefore, different structural features relevant to selectivity could be readily identified and the information might be further utilized in designing highly selective kinase inhibitors. The flexibility of the computational approach outlined in this chapter makes it suitable for the analysis of other profiling experiments with the goal of identifying active compounds that exhibit substantial variation in their activity profiles against different target proteins.

# Conclusion

Medicinal chemists often have to perform the non-trivial task of identifying various structural determinants within compound sets that influence their bioactivity. Systematic analysis of pair-wise molecular similarities and potency differences also helps to deduce SAR trends and formulate rules in order to guide different compound design or optimization attempts. The activity landscape concept is often employed to rationalize SARs in three dimensions, the first two depicting the chemical space followed by the addition of activity information as the third dimension. Thus, an activity landscape represents a hypersurface combining chemical similarity and biological activity data that is very similar to geographical maps in its topology.

The generation of 3D activity landscape models for real ligand sets has been one of the objectives of this dissertation. The resulting landscape representations largely depart from the idealized versions. However, the global SAR characteristics of bioactive compound sets are well preserved and intuitively accessible using these 3D views. Since, the topology of the landscapes is greatly influenced by the choice of the molecular representation, 3D landscape modeling can also be utilized to study the magnitude of alterations in SAR features brought about by alternative chemical spaces. Moreover, SAR relevant features like activity cliffs of varying magnitudes can also be identified.

An important characteristic of 3D activity landscapes is that proximity between ligands in the 2D projection correlates with their structural relatedness. By contrast, in 2D graphical landscape models like NSGs, placement of compounds and subsequent clustering is based on layout algorithms and has no chemical relevance. Therefore, the various conceptual differences in generating NSGs and 3D activity landscape representations make their comparison infor-

mative. Indeed, simultaneous examination of the visualizations obtained by these conceptually different methodologies has provided useful insights regarding the complementary global and local SAR information content associated with compound sets.

A computational approach has been introduced to calculate feature probabilities for individual compounds on the basis of their various SAR feature frequencies. By using fuzzy thresholds to derive these conditional probabilities, the existing SAR features have been used to generate eight feature categories. It has been demonstrated that subsequent assignment of compounds to these categories aided in their differentiation in local SAR environments when SAR analysis was performed using graphical activity landscape representations.

Typically, while analyzing SARs, one accounts for systematic structural similarity and potency distribution within ligand data sets. However, for ligands active against a receptor, information pertaining to its mechanism of action is also considered relevant. Routine approaches to analyzing SAR do not distinguish between ligands with different mechanisms of action. Adaptation of existing SAR analysis-driven data structures to incorporate mechanism related data has also been addressed in this dissertation. Graph-based landscape representations like NSGs are well suited for large-scale computational analysis and visualization of SAR as they are easy to navigate and interpretable. Introduction of a new color scheme sufficiently modified this activity landscape model to allow SAR as well as mechanism of action related analysis. The clear outcome of using M-NSGs is that compound subsets with either mechanistic homogeneity or heterogeneity can be readily identified. In addition, subsequent close inspection of structurally related ligand communities that are mechanistically heterogeneous can help characterize structural determinants that are responsible for switching the mechanism of action or "mechanism hops".

An inherent property of activity landscape representations based on chemical similarity calculated using molecular fingerprints is their "black box" nature. Thus, substructures or R-groups associated with compound potency are not apparent without the inspection of compound 2D structures. MMP-based approaches are better suited for the direct determination of chemical modifications resulting in potency improvement. Therefore, in order to facilitate assessment

of substructure changes accompanied by mechanism hops, an MMP-derived data structure was suitably modified. Application of M-BMMSGs to exemplary receptor ligand sets also demonstrated the ability of this approach to resolve mechanism hop inducing chemical replacements within these ligands at multiple levels.

Exploring relationships between chemical structure and bioactivity is commonly carried out for compound sets that are active against single targets. However, it is often necessary to monitor SAR trends in ligand data sets with activity annotations against multiple members of a target family so that compound selectivity patterns and off-target effects may be identified. Due to the intrinsic difficulty in navigating these high dimensional spaces, design of activity landscape representations that simplify access to multi-target relevant SAR information has also been attempted. Using a SOM-based 2D chemical space projection, a multi-target landscape model was generated in which compounds were represented in terms of binned pair-wise target activity differences. Such an encoding of compounds aided in the identification of continuous as well as discontinuous regions in multi-dimensional activity spaces. Examination of compounds within these regions revealed chemical replacements that retained or altered compound selectivity profiles.

Profiling experiments to perform simultaneous testing of compound libraries against many targets, especially high profile targets like protein kinases, have recently experienced increasing interest. Such profiling studies have made major contributions to the growth of multi-target activity data. A novel activity landscape model was designed to analyze the high-dimensional data generated from one such publicly available kinase profiling study. Ability to handle bioactivity annotations for very large numbers of targets and at the same time deal with incomplete activity matrix are two important features of this newly described landscape model.

The data obtained from another kinase profiling experiment was utilized to examine the ability of structurally related inhibitors to distinguish between various therapeutically relevant kinases. The differentiation potential was used as a measure to quantify the differential activity of these compounds against kinase targets. Compound ranking prioritized six inhibitors with very high differen-

tiation potential. Variation in the differentiation potential among the nearest neighbors of these top ranked compounds was also investigated. In addition, chemical modifications that produced alterations in the inhibition profiles of these compounds could be identified.

In conclusion, this dissertation reports novel approaches for the design of 2D and 3D activity landscapes to facilitate SAR analysis and visualization.

# Additional Publications

Iyer P., Hu Y., Bajorath J. SAR monitoring of evolving compound data sets using activity landscapes. *J. Chem. Inf. Model.*, **2011**, 51, 532-540.

Namasivayam V., Iyer P., Bajorath J. Extraction of discontinuous structure-activity relationships from compound data sets through particle swarm optimization. *J. Chem. Inf. Model.*, **2011**, 51, 1545-1551.

Namasivayam V., Iyer P., Bajorath J. Exploring SAR continuity in the vicinity of activity cliffs. *Chem. Biol. Drug Des.*, **2012**, 79, 22-29.

Wassermann A. M., Dimova D., Iyer P., Bajorath J. Advances in computational medicinal chemistry - matched molecular pair analysis. *Drug Develop. Res.*, **2012**, 73, 518-527.

Ahmadi M., Vogt M., Iyer P., Bajorath J., Fröhlich H. Predicting potent compounds via model-based global optimization. *J. Chem. Inf. Model.*, **2013**, 53, 553-559.

Iyer P., Stumpfe D., Vogt M., Bajorath J., Maggiora G. M. Activity landscapes, information theory, and structure-activity relationships. *Mol. Inf.*, **2013**, 32, 421-430.

# Lebenslauf

## Lebenslauf

Preeti Ramesh Iyer
Vorgebirgsstr. 6
53111 Bonn

Geboren am 30. Juni 1982 in Chennai, Indien

2000-2003: Bachelor of Science, Bangalore University, Bangalore

2004-2005: Master of Life Science Informatics, Vellore Institute of Technology, Vellore

2006-2007: Junior Microbiologist, Anand Diagnostic Laboratory, Bangalore

2007-2008: Microbiologist, Jupiter Hospital, Thane

2008-2010: Master of Life Science Informatics, Bonn-Aachen International Center for Information Technology (B-IT), Universität Bonn

2010-2013: Doctoral studies in Computational Life Sciences, Bonn-Aachen International Center for Information Technology (B-IT), Universität Bonn

# Eidesstattliche Erklärung

An Eides statt versichere ich hiermit, dass ich die Dissertation "Multi-faceted Structure-Activity Relationship Analysis Using Graphical Representations" 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 nachstehend aufgeführten Stellen auszugsweise veröffentlicht worden ist:

Peltason L., Iyer P., Bajorath J. Rationalizing three-dimensional activity landscapes and the influence of molecular representations on landscape topology and formation of activity cliffs. *J. Chem. Inf. Model.*, **2010**, 50, 1021-1033.

Iyer P., Wawer M., Bajorath J. Comparison of two- and three-dimensional activity landscape representations for different compound data sets. *Med. Chem. Comm.*, **2011**, 2, 113-118.

Vogt M., Iyer P., Maggiora G. M., Bajorath J. Conditional probabilities of activity landscape features for individual compounds. *J. Chem. Inf. Model.*, **2013**, 53, 1602-1612.

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. *J. Chem. Inf. Model.*, **2011**, 51, 1281-1286.

Iyer P., Bajorath J. Mechanism-based bipartite matching molecular series graphs to identify structural modifications of receptor ligands that lead to mechanism

hopping. *Med. Chem. Comm.*, **2012**, 3, 441-448.

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.

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.

Dimova D., Iyer P., Vogt M.,Totzke F., Kubbutat M. H. G., Schächtele C., Laufer S., Bajorath J. Assessing the target differentiation potential of imidazole-based protein kinase inhibitors. *J. Med. Chem.*, **2012**, 55, 11067-11071.

---

Preeti Ramesh Iyer
October 2013
Bonn