JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Conditional Probabilities of Activity Landscape Features for Individual Compounds

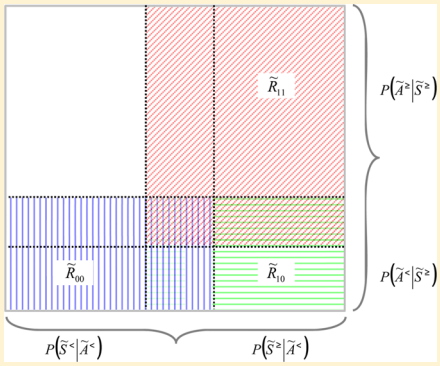
Martin Vogt,[†] Preeti Iyer,[†] Gerald M. Maggiora,*[,‡] and Jürgen Bajorath*[,†]

[†]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

[‡]College of Pharmacy & BIO5 Institute, University of Arizona, Translational Genomics Research Institute, 1295 North Martin, P.O. Box 210202, Tucson, Arizona 85721, United States, and 445 North Fifth Street, Phoenix, Arizona 85004, United States

Supporting Information

ABSTRACT: Activity landscape representations aid in the analysis of structure−activity relationships (SARs) of large compound data sets. Landscapes are characterized by features with different SAR information content such as, for example, regions formed by structurally diverse compounds having similar activity or, alternatively, structurally similar compounds with large activity differences, so-called activity cliffs. Modeling of activity landscapes typically requires pairwise comparisons of molecular similarity and potency relationships of compounds in a data set. Consequently, landscape features are generally resolved at the level of compound pairs. Herein, we introduce a methodology to assign feature probabilities to individual compounds. This makes it possible to organize compounds comprising activity landscapes into well-defined SAR categories. Specifically, the calculation of conditional feature probabilities of active compounds provides a balanced and further refined view of activity landscapes with a focus on individual molecules.



INTRODUCTION

Activity landscapes provide combined views of molecular similarity and potency relationships between active compounds and aid in structure−activity relationship (SAR) analysis of large data sets.[1] A variety of single- and multitarget activity landscape representations have been introduced, ranging from 2D plots and molecular network representations to 3D models.[1,2] Regardless of their details, the construction of activity landscapes typically requires systematic pairwise comparisons of compounds to determine their degree of structural relatedness and their potency differences. Consequently, activity landscape features are mostly discussed at the level of compound pairs. For example, activity cliffs, which are generally thought to be the most prominent features of activity landscapes, consist of pairs of structurally similar compounds with large potency differences.[3,4] In addition to activity cliffs other landscape features that are relevant for SAR analysis include pairs of compounds with similar structures and activity and pairs of compounds with different structures and similar activity, as further discussed below.

Although different activity landscape representations have been introduced, landscape features have mostly been described in qualitative terms. Recently, we have characterized activity landscape features in a more quantitative way from an information-theoretic perspective, compared their information content, and related the results of information entropy calculations to SAR information.[5] This study represented a first step to describe feature distributions in activity landscapes in a formal manner.

Herein, we develop a rigorous procedure based on concepts from probability theory to quantitatively analyze activity landscape features from a different, previously unconsidered point of view. From similarity and activity comparisons of compounds, we derive conditional probabilities for individual compounds to form activity cliffs and other predefined landscape features. Hence, the probability that any given compound in a data set will fall into different activity landscape regions is quantified.

In the following, we first describe the conceptual framework and derivation of our methodology in detail. Then, the methodology is applied to characterize a large number of data sets and identify compounds with significant feature probabilities. Finally, implications of our findings are discussed including practical applications of activity landscape modeling.

METHODOLOGY

Structure−Activity Similarity Maps. We initially classify activity landscape features using structure−activity similarity (SAS) maps,[6] the first 2D representation of activity landscapes. SAS maps resolve activity landscapes at the level of compound pairs and are particularly suited for classifying activity landscape features. For practical applications, several variants of SAS maps have also been introduced.[7,8]

Originally, SAS maps represented structural similarity and activity similarity of test compounds on the basis of pairwise

Received: May 15, 2013
Published: June 23, 2013

dx.doi.org/10.1021/ci400288r | J. Chem. Inf. Model. 2013, 53, 1602−1612

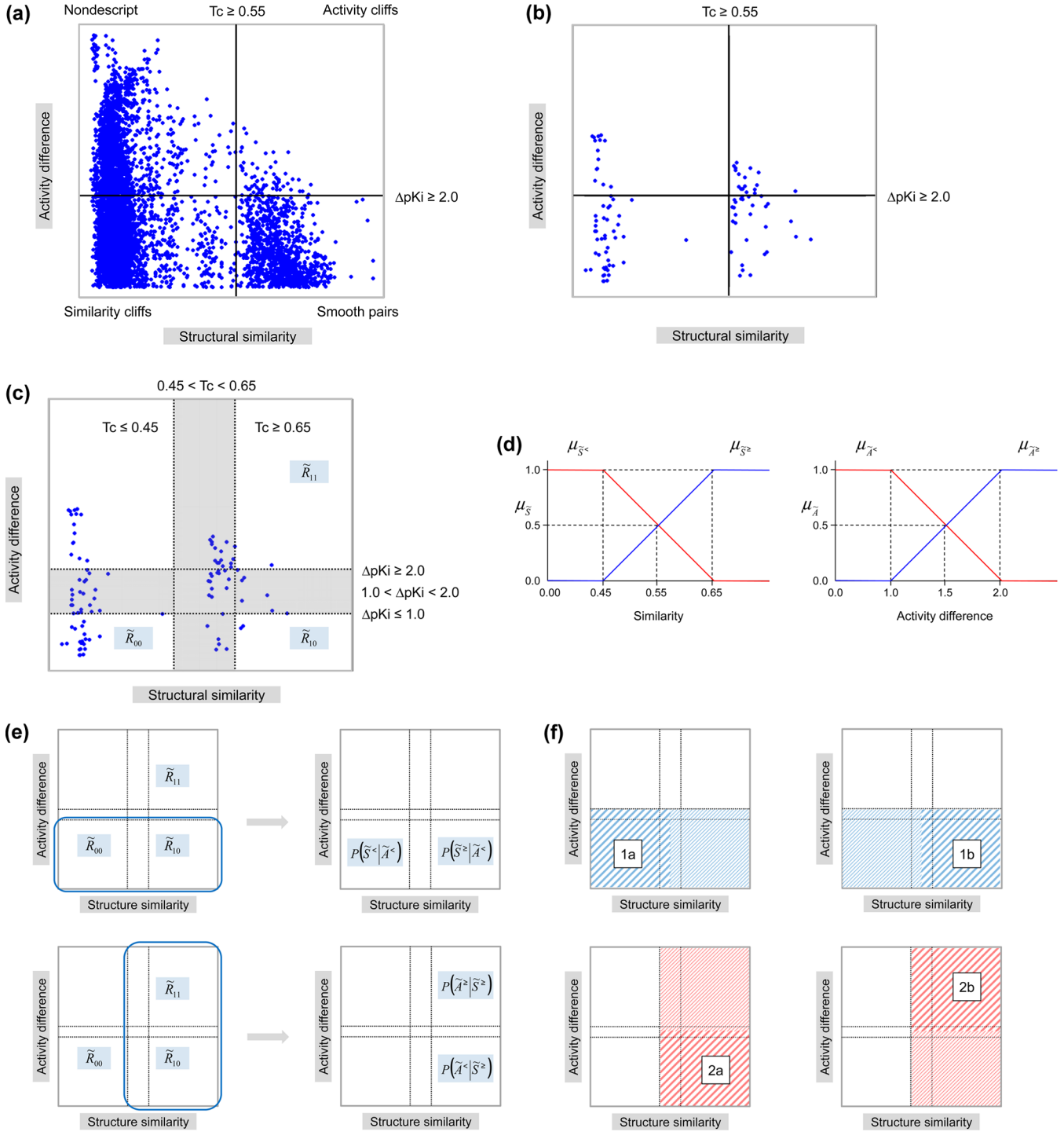


Figure 1. From activity landscape views to conditional per-compound feature probabilities. (**a**) For an exemplary data set, the structure−activity similarity (SAS) map representing all pairwise structural similarity and potency difference relationships is shown. The structural similarity and potency difference threshold values (Tc ≥ 0.55 and ΔpKi ≥ 2.0) are indicated by solid black lines separating the landscape into four distinct regions as indicated. (**b**) Local SAS map showing a subset of these relationships formed by an individual compound (i.e., this compound participates in each pair). (**c**) Crisp threshold values are replaced by fuzzy value intervals. (**d**) Crisp threshold values are replaced by "softer" boundaries exemplified by the membership functions of the fuzzy sets, $\mu_{\tilde{S}}$ and $\mu_{\tilde{A}}$, corresponding to similarities and absolute activity differences, respectively. (**e**) Derivation of conditional probabilities on the basis of local SAS maps under the condition that compound pairs have similar potencies (top) or are structurally similar (bottom). (**f**) Regions in the local SAS map are highlighted for the different conditional probability-based categories 1a, 1b, 2a, and 2b according to Table 1.

comparisons. Structural similarity is typically assessed by comparing compound fingerprints using the Tanimoto coefficient (Tc)[9] and activity similarity is determined by potency differences (normalized, if required) between pairs of compounds. No further parameters need to be taken into consideration at this stage. Each data point in the map represents a pairwise compound comparison. Structural similarity is reported on the *x*-axis and absolute potency difference on the *y*-axis, as depicted in

dx.doi.org/10.1021/ci400288r | J. Chem. Inf. Model. 2013, 53, 1602−1612

Figure 1a. As seen in the figure, this map can be subdivided into different regions based upon defined threshold values for structural similarity and potency difference. These regions represent different activity landscape features, as further detailed below.

Similarity and Potency Difference Thresholds. In order to carry out a statistical analysis of activity landscape features, threshold values for structural similarity and potency differences must be clearly defined.[1,4] For the purpose of our analysis, the similarity threshold is set to a Tc value of 0.55 (i.e., $s_T = 0.55$) calculated using extended connectivity fingerprints with bond diameter four (ECFP4),[10] and an absolute potency difference threshold of two pK_i units (i.e., $a_T = 2.00$), which is two log units of activity (i.e., 2 orders of magnitude - OoM). These threshold values have often been used to define activity cliffs.[4,11] Compound pairs with an ECFP4 Tc of 0.55 or higher are typically structurally very similar,[12] and an absolute potency difference of two OoM or larger accounts for significant activity cliffs in a data set.[11]

Activity Landscape Features. The threshold values delineate four different landscape feature regions in SAS maps, which are described as follows. The lower left section consists of structurally dissimilar pairs of compounds having similar potencies. These compound pairs, which might contain different scaffolds or have otherwise dissimilar structures, have previously been termed *similarity cliffs*,[5] a term that will also be used in this work. Compound pairs participating in similarity cliffs are typically well separated in chemical space. The lower right section of the map contains compound pairs with high structural similarity and small potency differences associated with smooth landscape regions and, hence, are termed *smooth pairs*. The upper right section contains structurally similar compound pairs with large potency differences, i.e., *activity cliffs*. Because of their high similarities, compound pairs in both of these classes are typically neighbors in chemical space. Lastly, the upper left region contains pairs of structurally dissimilar compounds with large potency differences typically found in relatively nondescript regions of activity landscapes. Since such compound pairs exhibit little if any useful SAR information, we focus predominantly on *similarity cliffs*, *activity cliffs*, and *smooth pairs* as primary activity landscape features.

Previously, we have applied information entropy to characterize the activity landscape features of SAS maps.[5] Information entropy calculations are based on frequency counts of compound pairs that lie in the different landscape regions of SAS maps. From an information theoretic point of view, activity cliffs, which are sparsely distributed, have high information content. Smooth pairs are more frequently observed and have lower information content. Interestingly, similarity cliffs, which were found to be prevalent in the activity landscapes of many data sets, formally have the lowest information content, although they often provide valuable information about structurally distinct compounds having similar activity.[5]

Per-Compound Feature Probabilities and Local SAS Maps. In this work, the focus of our methodology is on the landscape feature probabilities of individual molecules. *For any given compound in a data set, we determine its probability of participating in activity cliffs, similarity cliffs, or smooth pairs. Combinations of features are also considered.*

In support of this approach, *local SAS maps* are introduced that only show compound pairs formed by specific compounds. Each of these compound-based SAS maps reveals the propensity of a specific compound to form activity cliffs, similarity cliffs, or smooth pairs, as illustrated in Figure 1b. The information contained in a local SAS map can be calculated for any compound in the data set. For this purpose, we need to determine the frequencies with which a given compound, m_k, participates in each of the three different landscape feature regions of local SAS maps:

(I) Region $R_{00}(k)$ (lower left): Similarity Cliffs
(II) Region $R_{10}(k)$ (lower right): Smooth Pairs
(III) Region $R_{11}(k)$ (upper right): Activity Cliffs

The featureless region in the upper left corner of local SAS maps, designated by $R_{01}(k)$, is not considered further in this work (vide supra).

To clearly distinguish the two categories of SAS maps, maps that contain data on all compound pairs of a set of molecules will henceforth be designated as *global SAS maps* to contrast them with local SAS maps that are constructed with reference to a single compound, as described above.

Probabilities of Activity Landscape Features. Although the probability of activity landscape features can be calculated for compound pairs,[5] additional issues arise when the probabilities of the different landscape features are associated with specific compounds. Therefore, a more detailed description is given of the relevant aspects of probability theory applied to the present case. For clarity, this is initially formulated in terms of classical crisp sets, but this limitation is subsequently removed. In the following, the probabilistic foundation for a wide range of possible statistical analyses of activity landscape features is derived.[13]

The sample space Ω is used to describe all possible events that can occur. Given a set M of n compounds, the elementary events are described by compound pairs

$$\Omega = \{(m_{ij}) \in M \times M | 1 \le i, j \le n, i \ne j\} \tag{1}$$

where m_{ij} denotes the compound pair $(m_i m_j)$.[14] For finite sample spaces, the family of possible events is given by the set of all subsets, i.e. the power set, of the sample space $\mathcal{F} = \mathcal{P}(\Omega)$. A probability measure is then defined by its values on the elementary events

$$P(\{m_{ij}\}) = \frac{1}{|\Omega|} = \frac{1}{n(n-1)} \text{for all } 1 \le i, j \le n, i \ne j \tag{2}$$

where $|\Omega|$ is the cardinality (count of elements) of the sample space. Two random variables $S{:}\Omega \rightarrow \mathbf{R}$ and $A{:}\Omega \rightarrow \mathbf{R}$ are defined that reflect the *structural similarity* and *absolute potency difference*, respectively. Both functions are symmetric, i.e. $S(m_{ij}) = S(m_{ji})$ and $A(m_{ij}) = A(m_{ji})$.[15]

An SAS map can be seen as a projection $A{:}\Omega \rightarrow \mathbf{R}^2$ of the sample space onto the real plane by $R(m) = (S(m),A(m))$. Because of symmetry, both m_{ij} and m_{ji} will be projected onto the same point in $\mathbf{R}^2$. A local SAS map then corresponds to a subset $V_k \subset \Omega$ projected onto $\mathbf{R}^2$. The subsets of interest in this work are those in which one compound is fixed

$$V_k = \{m_{kj} \in \Omega | 1 \le j \le n, j \ne k\} \tag{3}$$

where $|V_k| = n-1$.

Applying random variables S and A for similarity and absolute activity difference and their corresponding threshold values s_T and a_T, respectively, events can be defined containing compound pairs with high or low structural similarity and/or absolute activity difference:[16]

$$S^{<} = \{m \in \mathbf{\Omega} | \mathrm{S}(m) < s_{\mathrm{T}}\}$$
$$S^{\geq} = \{m \in \mathbf{\Omega} | \mathrm{S}(m) \geq s_{\mathrm{T}}\}$$
$$A^{<} = \{m \in \mathbf{\Omega} | \mathrm{A}(m) < a_{\mathrm{T}}\}$$
$$A^{\geq} = \{m \in \mathbf{\Omega} | \mathrm{A}(m) \geq a_{\mathrm{T}}\} \quad (4)$$

Combining these events defines the four regions of local and global SAS maps described above. In the latter case, the crisp sets corresponding to the three regions of interest of a global SAS map are given by

$$R_{00} = \mathrm{R}^{-1}(\{(x, y) \in \mathbf{R}^2 | x < s_{\mathrm{T}}, y < a_{\mathrm{T}}\}) = S^{<} \cap A^{<} \text{ Similarity Cliffs}$$
$$R_{10} = \mathrm{R}^{-1}(\{(x, y) \in \mathbf{R}^2 | x \geq s_{\mathrm{T}}, y < a_{\mathrm{T}}\}) = S^{\geq} \cap A^{<} \text{ Smooth Pairs}$$
$$R_{11} = \mathrm{R}^{-1}(\{(x, y) \in \mathbf{R}^2 | x \geq s_{\mathrm{T}}, y \geq a_{\mathrm{T}}\}) = S^{\geq} \cap A^{\geq} \text{ Activity Cliffs} \quad (5)$$

Probabilities can now be determined for these composite events based on their respective cardinalities:

$$\Pr(\mathrm{S} < s_{\mathrm{T}}, \mathrm{A} < a_{\mathrm{T}}) = P(S^{<} \cap A^{<}) = \frac{|S^{<} \cap A^{<}|}{|\mathbf{\Omega}|} = \frac{|R_{00}|}{n(n-1)} \text{ Similarity Cliffs}$$
$$\Pr(\mathrm{S} \geq s_{\mathrm{T}}, \mathrm{A} < a_{\mathrm{T}}) = P(S^{\geq} \cap A^{<}) = \frac{|S^{\geq} \cap A^{<}|}{|\mathbf{\Omega}|} = \frac{|R_{10}|}{n(n-1)} \text{ Smooth Pairs}$$
$$\Pr(\mathrm{S} \geq s_{\mathrm{T}}, \mathrm{A} \geq a_{\mathrm{T}}) = P(S^{\geq} \cap A^{\geq}) = \frac{|S^{\geq} \cap A^{\geq}|}{|\mathbf{\Omega}|} = \frac{|R_{11}|}{n(n-1)} \text{ Activity Cliffs} \quad (6)$$

In our previous work,[5] the sample space, Ω', consisted of unique compound pairs; thus, m_{ij} and m_{ji} were not distinguished and $\Omega' = |\Omega|/2$. Note, however, that the cardinalities of all of the sets in the numerators are also exactly halved so that the calculated probabilities remain the same.

Using the above formalism, the probabilities associated with the activity landscape features of local SAS maps can be determined using an approach that is based on multiple conditioning. First, consider the local SAS maps. This corresponds to conditioning on one of the subsets V_k defined in eq 3, yielding the probabilities of the three landscape features, given that the first compound is m_k

$$\Pr(\mathrm{S} < s_{\mathrm{T}}, \mathrm{A} < a_{\mathrm{T}} | V_k) = P(S^{<} \cap A^{<} | V_k) = \frac{|S^{<} \cap A^{<} \cap V_k|}{|V_k|} = \frac{|R_{00}(k)|}{n-1} \text{ Similarity Cliffs}$$
$$\Pr(\mathrm{S} \geq s_{\mathrm{T}}, \mathrm{A} < a_{\mathrm{T}} | V_k) = P(S^{\geq} \cap A^{<} | V_k) = \frac{|S^{\geq} \cap A^{<} \cap V_k|}{|V_k|} = \frac{|R_{10}(k)|}{n-1} \text{ Smooth Pairs}$$
$$\Pr(\mathrm{S} \geq s_{\mathrm{T}}, \mathrm{A} \geq a_{\mathrm{T}} | V_k) = P(S^{\geq} \cap A^{\geq} | V_k) = \frac{|S^{\geq} \cap A^{\geq} \cap V_k|}{|V_k|} = \frac{|R_{11}(k)|}{n-1} \text{ Activity Cliffs} \quad (7)$$

where $R_{00}(k)$, $R_{10}(k)$, and $R_{11}(k)$ are defined as

$$R_{00}(k) = R_{00} \cap V_k = S^{<} \cap A^{<} \cap V_k$$
$$R_{10}(k) = R_{10} \cap V_k = S^{\geq} \cap A^{<} \cap V_k$$
$$R_{11}(k) = R_{11} \cap V_k = S^{\geq} \cap A^{\geq} \cap V_k \quad (8)$$

A second conditioning is employed because the ability of a given compound to form similarity cliffs should be evaluated only with respect to other compounds having similar activities, and, likewise, the ability of a compound to form activity cliffs should be judged only for those compounds that are structurally similar. Thus, the probabilities that compound pairs with similar activities (i.e., where $\mathrm{A} < a_{\mathrm{T}}$) will or will not form similarity cliffs are given by

$$\Pr(\mathrm{S} < s_{\mathrm{T}} | \mathrm{A} < a_{\mathrm{T}}, V_k) = P(S^{<} | A^{<}, V_k) = \frac{|S^{<} \cap A^{<} \cap V_k|}{|A^{<} \cap V_k|} = \frac{|S^{<} \cap A^{<} \cap V_k|}{|S^{<} \cap A^{<} \cap V_k| + |S^{\geq} \cap A^{<} \cap V_k|} = \frac{|R_{00}(k)|}{|R_{00}(k)| + |R_{10}(k)|} \quad (9a)$$

and

$$\Pr(\mathrm{S} \geq s_{\mathrm{T}} | \mathrm{A} < a_{\mathrm{T}}, V_k) = P(S^{\geq} | A^{<}, V_k) = \frac{|S^{\geq} \cap A^{<} \cap V_k|}{|A^{<} \cap V_k|} = \frac{|S^{\geq} \cap A^{<} \cap V_k|}{|S^{<} \cap A^{<} \cap V_k| + |S^{\geq} \cap A^{<} \cap V_k|} = \frac{|R_{10}(k)|}{|R_{00}(k)| + |R_{10}(k)|} \quad (9b)$$

It is clear from these equations that the probability is conserved, i.e. $\Pr(\mathrm{S} < s_{\mathrm{T}} | \mathrm{A} < a_{\mathrm{T}}, V_k) + \Pr(\mathrm{S} \geq s_{\mathrm{T}} | \mathrm{A} < a_{\mathrm{T}}, V_k) = 1$.

dx.doi.org/10.1021/ci400288r | J. Chem. Inf. Model. 2013, 53, 1602−1612

In the second case, the probabilities that compound pairs of structurally similar molecules (i.e., where $S \geq s_T$) will not or will form activity cliffs are given by

$$\Pr(A < a_T | S \geq s_T, V_k) = P(A^< | S^\geq, V_k) = \frac{|A^< \cap S^\geq \cap V_k|}{|S^\geq \cap V_k|} = \frac{|A^< \cap S^\geq \cap V_k|}{|A^< \cap S^\geq \cap V_k| + |A^\geq \cap S^\geq \cap V_k|} = \frac{|R_{10}(k)|}{|R_{10}(k)| + |R_{11}(k)|} \quad (10a)$$

and

$$\Pr(A \geq a_T | S \geq s_T, V_k) = P(A^\geq | S^\geq, V_k) = \frac{|A^\geq \cap S^\geq \cap V_k|}{|S^\geq \cap V_k|} = \frac{|A^\geq \cap S^\geq \cap V_k|}{|A^< \cap S^\geq \cap V_k| + |A^\geq \cap S^\geq \cap V_k|} = \frac{|R_{11}(k)|}{|R_{10}(k)| + |R_{11}(k)|} \quad (10b)$$

where, as before, probability is conserved, i.e. $\Pr(A < a_T | S \geq s_T, V_k) + \Pr(A \geq a_T | S \geq s_T, V_k) = 1$

Crisp versus Fuzzy Boundaries. Compound pairs can be assigned to any one of the three regions depending on their similarity values and the differences in their activities. However, using the *crisp* thresholds described earlier, assignment of a compound pair to a particular map region can be prone to boundary effects, especially for small sample sizes. For example, minute variations of the similarity threshold value could determine whether a given compound pair is classified as an activity cliff or as a nondescript pair. In an analogous fashion, small variations in potency differences could determine whether pairs of compounds are classified as activity cliffs or smooth pairs. Such dramatic changes in classification for small differences about the threshold values are not chemically meaningful. Therefore, in order to reduce boundary effects that could negatively affect feature assignments, we have replaced *crisp* thresholds with *fuzzy* boundaries ("*twilight zones*"), shown shaded in gray in Figure 1c. The fuzzy boundaries transform the crisp sets described above into the fuzzy sets depicted in Figure 1d.[17,18] The ordinates in the figure correspond to membership functions $\mu_{\tilde{S}}(m_{ij})$ and $\mu_{\tilde{A}}(m_{ij})$ of the fuzzy sets $\tilde{S}$ and $\tilde{A}$ associated with similarities and absolute activity differences, respectively. The tilde "~" above a set symbol indicates that the respective set is fuzzy. The membership function values, which lie on the unit interval of the real line $[0,1]$, describe the degree to which a compound pair belongs to a given fuzzy set. The complements of the two fuzzy sets are, respectively,

$$\mu_{\tilde{X}^\geq}(m_{ij}) = 1 - \mu_{\tilde{X}^<}(m_{ij}) \text{ for } 1 \leq i, j \leq n, i \neq j \quad (11)$$

where $\tilde{X} = \tilde{S}$ or $\tilde{A}$.

It is clear from eq 11 and from Figure 1d that for a given compound pair the sum of the membership values for each of the fuzzy sets and their complements is equal to unity. Because of this, the crisp partitions of local or global SAS maps described earlier are transformed into fuzzy partitions. However, because of eq 11 cardinalities are preserved, i.e.

$$|\tilde{X}^<| + |\tilde{X}^\geq| = \sum_{i,j=1,i\neq j}^{n} \mu_{\tilde{X}^<}(m_{ij}) + \sum_{i,j=1,i\neq j}^{n} \mu_{\tilde{X}^\geq}(m_{ij}) = \sum_{i,j=1,i\neq j}^{n} (\mu_{\tilde{X}^<}(m_{ij}) + \mu_{\tilde{X}^\geq}(m_{ij})) = \sum_{i,j=1,i\neq j}^{n} 1 = n(n-1) = |\Omega| \quad (12)$$

where $\tilde{X} = \tilde{S}$ or $\tilde{A}$. Thus, this approach, which is based on what can now be called "fuzzy landscape features" makes it possible to adhere to the original data partitioning scheme of local SAS maps while softening the boundaries between the different regions.

The mathematical expressions that define the membership functions as depicted in Figure 1e are given in eqs 13a and 13b for similarities and in eqs 14a and 14b for absolute activity differences:

$$\mu_{\tilde{S}^<}(m_{ij}) = \begin{cases} 1 & \text{if } Tc(m_{ij}) \leq 0.45 \\ \frac{1}{0.2}(0.65 - Tc(m_{ij})) & \text{if } 0.45 < Tc(m_{ij}) < 0.65 \\ 0 & \text{if } 0.65 \leq Tc(m_{ij}) \end{cases} \quad (13a)$$

$$\mu_{\tilde{S}^\geq}(m_{ij}) = \begin{cases} 0 & \text{if } Tc(m_{ij}) \leq 0.45 \\ \frac{1}{0.2}(Tc(m_{ij}) - 0.45) & \text{if } 0.45 < Tc(m_{ij}) < 0.65 \\ 1 & \text{if } 0.65 \leq Tc(m_{ij}) \end{cases} \quad (13b)$$

$$\mu_{\tilde{A}^<}(m_{ij}) = \begin{cases} 1 & \text{if } \Delta pKi(m_{ij}) \leq 1 \\ 2 - \Delta pKi(m_{ij}) & \text{if } 1 < \Delta pKi(m_{ij}) < 2 \\ 0 & \text{if } 2 \leq \Delta pKi(m_{ij}) \end{cases} \quad (14a)$$

$$\mu_{\tilde{A}^\geq}(m_{ij}) = \begin{cases} 0 & \text{if } \Delta pKi(m_{ij}) \leq 1 \\ \Delta pKi(m_{ij}) - 1 & \text{if } 1 < \Delta pKi(m_{ij}) < 2 \\ 1 & \text{if } 2 \leq \Delta pKi(m_{ij}) \end{cases} \quad (14b)$$

Importantly, as expressed by eqs 11 and 12, these membership functions conserve the cardinality of Ω and thus, form a proper foundation for computing probabilities.

Composite membership functions can be obtained for each of the four landscape features that lie in regions $\tilde{R}_{00}(k)$, $\tilde{R}_{10}(k)$, $\tilde{R}_{11}(k)$, and $\tilde{R}_{01}(k)$ of local SAS maps by taking intersections of appropriate pairs of fuzzy subsets.[19] In set-theoretic terms, such composite membership functions correspond to binary relations generated by the *intersection* of the fuzzy sets, i.e. $\tilde{S}^< \cap \tilde{A}^<$. For example, $\tilde{R}_{00}(k)$, which corresponds to similarity cliffs, contains compound pairs with low potency difference and low structural similarity.

Unlike the case for classical, crisp sets, where there is a single definition for the intersection of two sets, this is not the case for fuzzy sets, where set intersection can be defined in different ways. For fuzzy sets, the concept of set intersection is typically defined in terms of membership functions, where it is common to take the *minimum* of the membership function values corresponding to appropriate pairs of elements in the sets being intersected.[20] Here, *algebraic intersection*, which is obtained by multiplying the memberships functions, is used. Thus, the membership functions for fuzzy relations corresponding to the four regions of local SAS maps are given by

dx.doi.org/10.1021/ci400288r | J. Chem. Inf. Model. 2013, 53, 1602−1612

$$\mu_{\tilde{R}_{00}}(m_{ij}) = \mu_{\tilde{S}^{<}}(m_{ij})\cdot\mu_{\tilde{A}^{<}}(m_{ij}) \quad \text{Similarity Cliffs}$$
$$\mu_{\tilde{R}_{10}}(m_{ij}) = \mu_{\tilde{S}^{\geq}}(m_{ij})\cdot\mu_{\tilde{A}^{<}}(m_{ij}) \quad \text{Smooth Pairs}$$
$$\mu_{\tilde{R}_{11}}(m_{ij}) = \mu_{\tilde{S}^{\geq}}(m_{ij})\cdot\mu_{\tilde{A}^{\geq}}(m_{ij}) \quad \text{Activity Cliffs}$$
$$\mu_{\tilde{R}_{01}}(m_{ij}) = \mu_{\tilde{S}^{<}}(m_{ij})\cdot\mu_{\tilde{A}^{\geq}}(m_{ij}) \quad \text{Nondescript} \tag{15}$$

Because of the fuzzy boundaries (twilight zones) associated with each of the four classes, a given compound pair may belong to more than one class. Due to the complementary nature of the pairs of fuzzy sets $(\tilde{S}^{<},\tilde{S}^{\geq})$ and $(\tilde{A}^{<},\tilde{A}^{\geq})$ (see eq 11) the total membership over all four regions for any compound pair can be to shown to be equal to unity, i.e.

$$\mu_{\tilde{R}_{00}}(m_{ij}) + \mu_{\tilde{R}_{01}}(m_{ij}) + \mu_{\tilde{R}_{10}}(m_{ij}) + \mu_{\tilde{R}_{11}}(m_{ij}) = 1 \text{ for all } m_{ij} \in \mathbf{\Omega} \tag{16}$$

As a result, the algebraic set intersection procedure has a crucial advantage for this work over the usual "minimum" set-intersection procedure, since use of the latter does not conserve overall membership in the set of fuzzy relations (Cf. eq 16).

Because membership in the nondescript class of landscape features is neglected in this work, the sum of the membership function values for any given compound pair is less than unity in all but extremely rare cases. However, as will be seen in the sequel, this does not present any problems because conditional probabilities are used and thus, information on the complete sample space is not required (see e.g. eq 7).

Frequencies of Fuzzy Landscape Features. In each of three regions of interest in this work, the total frequency of compound pairs associated with the k-th reference compound is given by

$$|\tilde{R}_{00}(k)| = \sum_{j\neq k} \mu_{\tilde{R}_{00}(k)}(m_{kj})$$
$$|\tilde{R}_{10}(k)| = \sum_{j\neq k} \mu_{\tilde{R}_{10}(k)}(m_{kj})$$
$$|\tilde{R}_{11}(k)| = \sum_{j\neq k} \mu_{\tilde{R}_{11}(k)}(m_{kj}) \tag{17}$$

It should be noted that the corresponding frequencies might be fractional, given that partial memberships within twilight zones yield values of less than one. These frequencies can then be used to determine the probability that a given compound will form activity cliffs, similarity cliffs, or smooth pairs. Because the basic features of local and global SAS maps are preserved when fuzzy sets are substituted for crisp sets, one can refer to the corresponding SAS maps as *fuzzy local or fuzzy global SAS maps*.

Conditional Probabilities of Fuzzy Landscape Features. As was discussed above, the fuzzy partitioning of activity landscape features preserves all of the probabilistic characteristics ascribed to crisp sets, hence simply inserting fuzzy sets for their corresponding crisp sets appropriately modifies eqs 9 and 10. Equations 9a and 9b become

$$P(\tilde{S}^{<}|\tilde{A}^{<}, V_k) = \frac{|\tilde{R}_{00}(k)|}{|\tilde{R}_{00}(k)| + |\tilde{R}_{10}(k)|} \tag{18a}$$

$$P(\tilde{S}^{\geq}|\tilde{A}^{<}, V_k) = \frac{|\tilde{R}_{10}(k)|}{|\tilde{R}_{00}(k)| + |\tilde{R}_{10}(k)|} \tag{18b}$$

while eqs 10a and 10b become

$$P(\tilde{A}^{<}|\tilde{S}^{\geq}, V_k) = \frac{|\tilde{R}_{10}(k)|}{|\tilde{R}_{10}(k)| + |\tilde{R}_{11}(k)|} \tag{19a}$$

$$P(\tilde{A}^{\geq}|\tilde{S}^{\geq}, V_k) = \frac{|\tilde{R}_{11}(k)|}{|\tilde{R}_{10}(k)| + |\tilde{R}_{11}(k)|} \tag{19b}$$

Note that V_k remains a crisp set since its elements are the compound pairs associated with m_k. As was the case for conditional probabilities of crisp sets, the probabilities associated with the fuzzy sets are also conserved.

Using the conditional probabilities in eqs 18 and 19 it is possible to classify compound pairs into the following categories based on their probability of forming specific landscape features:

1. Two compounds with similar potencies
   a. with a probability for forming a similarity cliff (eq 18a)
   b. with a probability for forming a smooth pair (eq 18b)
2. Two structurally similar compounds
   a. with a probability to form a smooth pair (eq 19a)
   b. with a probability to form an activity cliff (eq 19b)

It should be noted, however, that although probabilities given by eqs 18b and 19a account for the probability to form a smooth pair, they are distinct. In the first case the probability is calculated for all pairs with similar potency differences, whereas in the second case it is calculated for all structurally similar pairs.

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 conditional probabilities estimated by eqs 19a and 19b if a compound had only very few structural neighbors. Therefore, only probability calculations based upon denominators $(|\tilde{R}_{00}(k)| + |\tilde{R}_{10}(k)|$ and $|\tilde{R}_{10}(k)| + |\tilde{R}_{11}(k)|,)$ greater than 2.0 were considered.

In order to identify significant probabilities, all probabilities calculated for a representative collection of data sets were sorted and probability thresholds, $P_T(\tilde{S}^{<}|\tilde{A}^{<},V_k)$, $P_T(\tilde{S}^{\geq}|\tilde{A}^{<},V_k)$, $P_T(\tilde{A}^{<}|\tilde{S}^{\geq},V_k)$, and $P_T(\tilde{A}^{\geq}|\tilde{S}^{\geq},V_k)$, were determined at the 90th percentile for the corresponding probabilities given in eqs 18 and 19. Probabilities above these values were considered significant.

Assessing the significance of the respective conditional probabilities with respect to a large collection of data sets allows the identification of exceptional probabilities irrespective of the absolute magnitude of these probabilities. For example, activity cliffs represent a generally rare activity landscape feature. The majority of compounds in a data set are not expected to form activity cliffs with their immediate structural neighbors. Comparably low probabilities might be significant for activity cliffs, but these probabilities would be much lower than those of similarity cliffs, which typically dominate activity landscapes.

Furthermore, the use of conditional feature probabilities compared to absolute probabilities offers conceptual advantages. For example, the conditional probability of a compound to form activity cliffs only takes account of structurally similar compounds and is thus independent of dissimilar compounds in the data set. On the other hand, absolute probabilities are highly influenced by the number of dissimilar compounds and therefore are difficult to interpret and only poorly reflect the

ability of a compound to form activity cliffs. In a similar fashion, the conditional probability of compounds to form similarity cliffs only takes compounds with similar potencies into account. For instance, if we consider a structurally diverse data set with few highly potent and many weakly potent compounds, then a weakly potent compound would be expected to participate in similarity cliffs more frequently than a highly potent compound simply because the number of weakly potent compounds is larger. The conditional probability takes this bias into account by relating the number of similarity cliff-forming pairs for a specific compound to the number of all compounds with similar potencies, thus yielding comparable values for the potential to form similarity cliffs.

Refined Activity Landscape Features. On the basis of conditional probabilities, the analysis of activity landscape features can be further refined. Because a compound can have a high probability for either category 1a or 1b and also for either category 2a or 2b, eight feature categories can be defined for SAR-relevant activity landscape regions, as reported in Table 1 (categories 1−8). Regions corresponding to categories 1−4 are delineated in Figure 1f. On the basis of conditional probabilities, we can assign the probability that a compound will form activity cliffs, similarity cliffs, smooth pairs, or some combination of these features. For example, category 1a in Table 1 focuses on the similarity cliff region in a local SAS map and characterizes compounds that possess a high probability to form similarity cliffs with other compounds that have similar potencies. The significance criterion is met if the corresponding conditional probability exceeds the threshold. Analogously, compounds in category 2b have a significant probability to form activity cliffs, given that they are structurally similar to others. Compounds that are likely to form smooth pairs can be further differentiated with respect to their probability to either form similarity cliffs (category 1b) or activity cliffs (category 2a), as also illustrated in Figure 1f. In addition, we can account for the probabilities of compounds to form four other types of feature combinations, given that a compound has similar activity to others or that it is structurally similar to others. For example, combined category 1a−2a accounts for the case that a compound is likely to form similarity cliffs and unlikely to form activity cliffs, whereas category 1b−2b accounts for the opposite case, i.e., a compound is likely to form activity cliffs but unlikely to form similarity cliffs. These combinations and the corresponding per-compound probabilities further refine our current views of activity landscape features and their distribution.

For each compound, the four conditional probabilities are calculated on the basis of full or partial memberships (due to the use of fuzzy boundaries) to different regions in local SAS maps.

Table 1. Activity Landscape Feature Probability Classification[a]

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

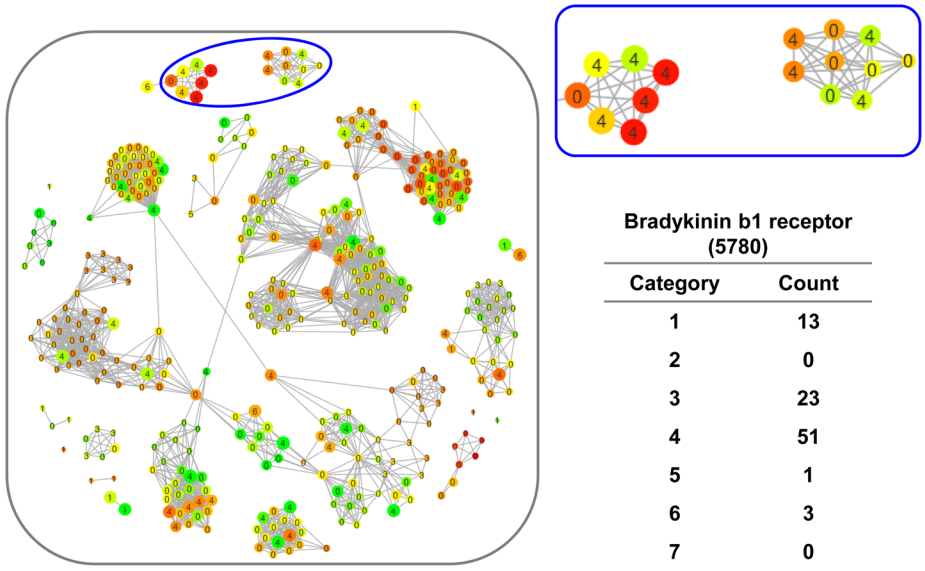
[a]Different activity landscape feature categories (cat) are reported. Categories consist of single features or combinations of features (type). For each category, the significance criterion is given and the corresponding landscape feature probabilities are described.

APPLICATIONS

Feature Probabilities and Thresholds. Conditional feature probabilities were calculated for 139 different activity classes extracted from BindingDB.[21] The thresholds $P_T(\tilde{S}^{<}|\tilde{A}^{<},V_k)$, $P_T(\tilde{S}^{\geq}|\tilde{A}^{<},V_k)$, $P_T(\tilde{A}^{<}|\tilde{S}^{\geq},V_k)$, and $P_T(\tilde{A}^{\geq}|\tilde{S}^{\geq},V_k)$ (see the Methodology section) were determined on the basis of the combined conditional probabilities from all 139 classes.

Compound Assignment. Based on the criteria given in Table 1, each compound was assigned to a single category corresponding to the conditional probability or combinations of conditional probabilities that exceeded the respective thresholds given by the significance criterion. Some of the categories in Table 1 are mutually exclusive. For example, a compound cannot possibly be assigned to categories 2a and 2b (i.e., *smooth pairs likely/activity cliffs unlikely* vs *activity cliff likely*). Other categories are not exclusive. For example, a compound can be assigned to categories 2b and 1a,2b (i.e., *activity cliffs likely* vs *similarity cliffs likely/activity cliffs likely*). If a compound exceeds the threshold value for a single feature and also for a combination involving this feature, it was assigned to the combination.

Visualization. SAS maps can be elegantly used to conceptualize and explain activity landscape features, hence providing a basis for our analysis, but they are not suitable for graphically analyzing per-compound feature probabilities because their basic data points (units) are compound pairs. In order to visualize feature probabilities and compound assignments, a compound network-based activity landscape representation was utilized that resolves the landscape at the level of individual compounds rather than pairs. In the Network-like Similarity Graph (NSG),[22] depicted in Figure 2, nodes represent compounds that are color-coded according to their potency values (from green/low potency over yellow to red/high potency) and edges represent similarity relationships (i.e., two compounds are connected by an edge if their calculated ECFP4 Tc value is equal to or greater than 0.55). Furthermore, nodes are scaled in size according to compound discontinuity scores.[22,23] Thus, the larger the node, the larger the degree of discontinuity a compound introduces in the activity landscape.



Bradykinin b1 receptor (5780)

Category	Count
1	13
2	0
3	23
4	51
5	1
6	3
7	0
8	0

Figure 2. Visualization of categorized compounds. A Network-like Similarity Graph (NSG) is shown for an exemplary data set. Individual compounds are represented as nodes and edges indicate pairwise similarity relationships. Nodes are labeled with their conditional probability-based categories according to Table 1. A section of the NSG is encircled and enlarged on the right. The numbers of compounds belonging to the different feature categories along with the target name and an in-house assigned target ID is reported in the table insert.

Activity cliffs represent the extreme form of SAR discontinuity.[4] Hence, in an NSG, combinations of large red and green nodes connected by edges indicate the most prominent activity cliffs present in a data set. In NSGs (and also for the calculation of discontinuity scores), a crisp similarity threshold is applied to establish pairwise compound similarity realtionships. It should also be noted that the 2D arrangement of compounds and clusters in an NSG has essentially no chemical meaning. Rather, relative compound and cluster positions and the distances separating them are determined by a graphical layout algorithm that forms densely connected clusters of similar compounds and separates them for clarity. For feature probability display, we annotate nodes with the assigned category according to Table 1. In Figure 2, a complete NSG for an exemplary data set is shown, and a highlighted section of the NSG is further enlarged for a more detailed inspection. This global vs local representation scheme is utilized in the following.

Systematic Calculations and Representative Examples. When systematically deriving feature probabilities for individual compounds comprising all 139 data sets, we observed significant differences in feature probability distributions. A number of data sets were identified that contained compounds belonging to diverse feature categories, without obvious preferences, whereas other data sets were dominated by compounds belonging to one or two feature categories. In 27 data sets, varying numbers of compounds belonging to categories 1−4 were found but no compounds representing combinations of conditional probabilities (categories 5−8). In addition, we also observed that data sets typically contained varying numbers of insignificant compounds (category 0) that were not likely to yield interpretable SAR information.

In Figure 3, we provide selected data sets focusing on each of the eight feature categories (1−8) according to Table 1 and report the per-compound conditional feature probability distributions. These examples illustrate the variety of distributions we observed and focus on individual feature categories. The complete NSGs for these exemplary compound sets are provided in Figure S1 of the Supporting Information. In Figure 3, selected sections of these NSGs are shown that contain compounds with specific feature probabilities. On the basis of such graph representations in combination with feature category information, medicinal chemistry hypotheses can often be developed. For example, subsets of compounds having a high probability to form activity cliffs can be selected and chemical changes inspected that are likely to represent activity determinants. Alternatively, scaffolds hops can be identified by focusing on compounds having a high probability to form similarity cliffs. Furthermore, during hit-to-lead and lead optimization studies, feature categories might be helpful to compare ligands directed against families of targets such as kinases. In this case, feature categories of individual compounds might be compared across different targets to identify compounds having the largest potential to yield highly potent inhibitors for a given target. For example, high probability of a compound to form activity cliffs for a given target might make this compound an attractive candidate for chemical exploration if it displays different probabilities for other targets.

The serotonin receptor-7 ligand set in Figure 3a is dominated by compounds having a high probability for forming similarity cliffs (category 1). Many of these compounds have intermediate potency and only limited structural similarity to others. In addition, the data set also contains many insignificant compounds (failing to reach the threshold for any of the eight SAR-relevant categories). Thus, this structurally diverse and heterogeneous data set has only limited SAR information content. The selected subset also contains an insignificant compound (category 0). By contrast, in the endothelin receptor ligand set in Figure 3b, the majority of compounds belong to category 2, i.e., these compounds are likely to form smooth pairs and unlikely to form similarity cliffs. Consistent with these observations, the NSG of the data set is dominated by densely connected clusters of structurally similar compounds with similar potency (Figure S1b of the Supporting Information).

dx.doi.org/10.1021/ci400288r | J. Chem. Inf. Model. 2013, 53, 1602−1612

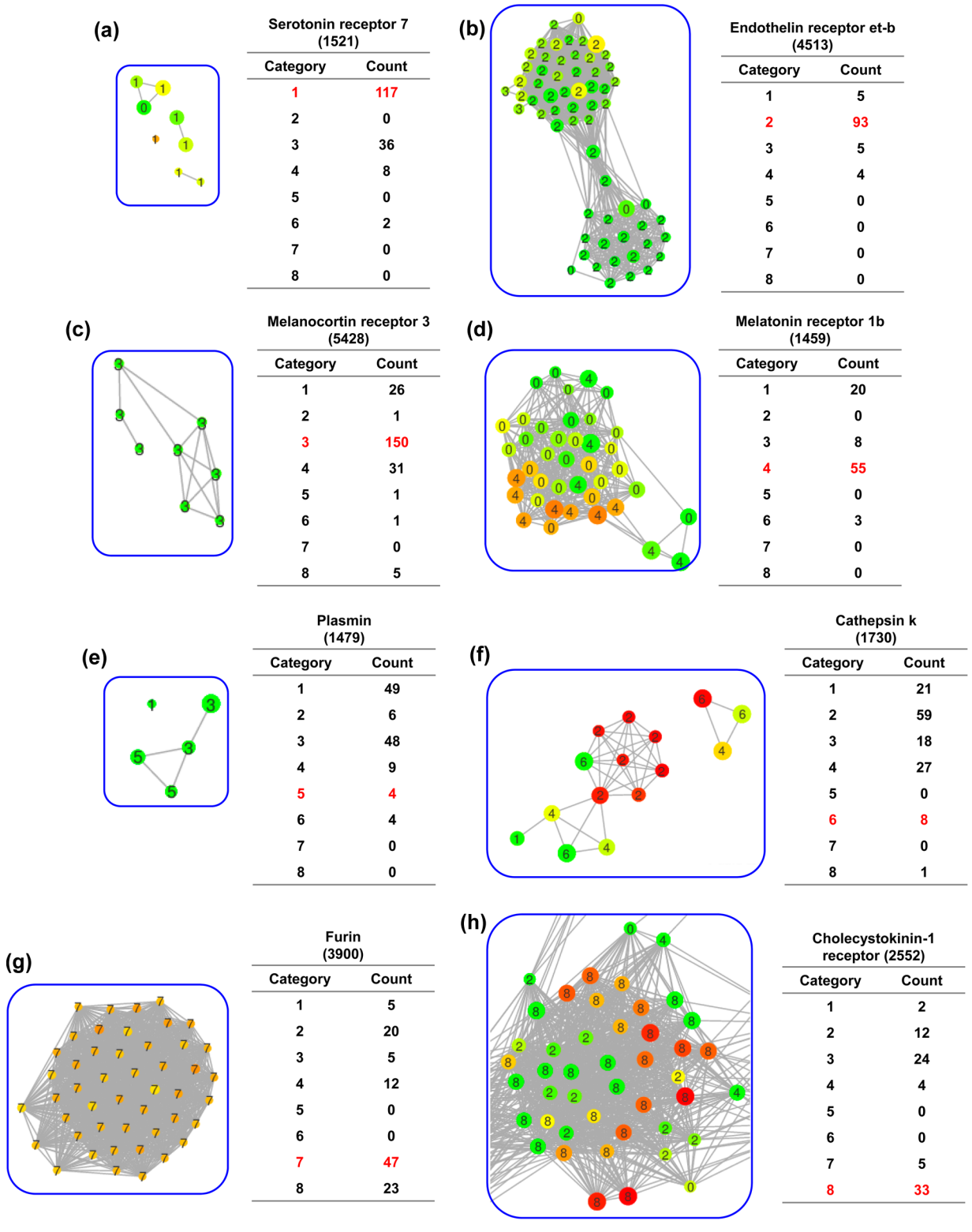


Figure 3. Representative compound subsets. In (**a**)−(**h**), enlarged sections of NSGs are shown for eight different compound data sets according to Figure 2. The complete NSGs from which the sections are taken are shown in Figure S1 of the Supporting Information. Each representation focuses on an NSG region containing compounds belonging to a feature category frequently observed in the data set (highlighted in red in the table inserts). (**a**) Data set 1521, (**b**) 4513, (**c**) 5428, (**d**) 1459, (**e**) 1479, (**f**) 1730, (**g**) 3900, (**h**) 2552.

Similarly, in the melanocortin receptor 3 ligand data set in Figure 3c, many compounds are likely to form smooth pairs but unlikely to form activity cliffs (category 3). The complete NSG of this data set (Figure S1c of the Supporting Information) nicely illustrates the presence of structurally distinct subsets of compounds with different SAR information content that are separated from each other.

The melatonin receptor 1b ligand set in Figure 3d contains 55 compounds with a high probability to form activity cliffs (category 4). In the displayed densely connected subset, com-

dx.doi.org/10.1021/ci400288r | J. Chem. Inf. Model. 2013, 53, 1602−1612

binations of large red and green nodes are apparent that indicate prominent activity cliffs in the data set. Interestingly, this cluster also contains many insignificant compounds, which mostly have intermediate potency and smaller node radii; i.e., they introduce only little SAR discontinuity in the activity landscape, despite their apparent association with cliffs. On the basis of the graphical representations compounds in categories 4 and 0 essentially cannot be distinguished from each other. However, taking conditional feature probabilities into account, compounds that are most likely to form activity cliffs can de differentiated from others, which further refines the activity landscape view.

In the plasmin inhibitor set in Figure 3e, compounds that are likely to form similarity cliffs and unlikely to form activity cliffs (category 5) are emphasized. A small compound subset is shown where these category 5 compounds are connected to compounds that are likely to form smooth pairs and also unlikely to form activity cliffs (category 3). Again, in the graphical representation, these compounds are equivalent and cannot be further differentiated without information from conditional probability.

The cathepsin K inhibitor set in Figure 3f contains eight compounds belonging to category 6, i.e., these compounds are likely to form both similarity cliffs and activity cliffs. The selected compound subset includes four of these compounds, three of which have relatively low potency. Two of these compounds form an activity cliff but are structurally distinct from the other two. The weakly potent cliff partner also participates in similarity cliffs with the remaining three compounds. *This information cannot be deduced from the NSG representation.* The two activity cliff compounds are also connected to a category 4 compound that is also likely to form activity cliffs. In addition, two of the category 6 compounds participate in activity cliffs with compounds having a high probability to form smooth pairs (category 2), i.e., structurally similar compounds with comparable potency. With one exception, all of these category 2 compounds are only involved in one activity cliff but in six smooth pairs each. Thus, taken together, this compound cluster represents a highly differentiated local SAR environment on the basis of conditional probabilities.

The furin inhibitor set in Figure 3g contains 47 compounds with low similarity cliff and low activity cliff probability. The displayed densely connected cluster contains all of these compounds that have very similar potencies. This subset of compounds structurally differs from the remainder of the data set and forms smooth pairs. Because these compounds yield high probabilities for both categories 1b and 2a, they are ultimately assigned to category 7 representing type 1b−2a.

Finally, in Figure 3h, a cluster of cholecystokinin-1 receptor ligands is shown that dominates this data set and contains many category 8 compounds that are unlikely to form similarity cliffs but likely to form activity cliffs. In addition, this cluster also contains small numbers of compounds belonging to categories 0, 2, and 4. Many category 8 compounds are involved in the formation of multiple and large-magnitude activity cliffs. The two category 4 compounds in this cluster with a high probability for forming activity cliffs are weakly potent and participate in cliffs with category 8 compounds. Overall, this cluster is rich in SAR information as reflected by the presence of many large red and green nodes and the prevalence of category 8 compounds.

Taken together, the examples in Figure 3 and Figure S1 of the Supporting Information illustrate that conditional landscape feature probabilities help to further differentiate between active compounds in activity landscapes and the SAR information they are associated with.

CONCLUSIONS

We have introduced a new methodology for assigning activity landscape features to individual compounds. The approach is based on the determination of conditional feature probabilities using fuzzy boundaries between different activity landscape regions. This makes it possible to consider compounds with partial memberships in multiple landscape features and provides a balance to possible boundary effects, which we have found to be especially useful for analyzing compound data sets of limited or moderate size. Conditional probabilities are derived from pairwise compound similarity and potency comparisons, i.e., basic data for activity landscape generation, by focusing feature frequency analysis on each compound in a data set. For this purpose, local SAS maps have been introduced. The ensuing assignment of conditional probabilities to individual compounds provides a conceptual advance in activity landscape analysis. This is reflected by the ability to categorize compounds and further differentiate between compounds in local SAR environments that are difficult to deconvolute on the basis of graphical analysis, even taking numerical SAR discontinuity measures into account. Feature probabilities are descriptive in nature and can therefore not directly be used to predict the activity of new compounds. The use of conditional probabilities has made it possible to derive eight different feature categories from three primary SAR-relevant categories of compound pairs including activity cliffs, similarity cliffs, and smooth pairs. This categorization scheme further refines current activity landscape views and supports systematic SAR analysis at the level of individual compounds.

ASSOCIATED CONTENT

Supporting Information

Supplementary Table S1 lists all compound data sets that were analyzed. Supplementary Figure S1 shows complete NSG representations for the compound data sets discussed in the text. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Phone: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de (J.B.), gerry.maggiora@gmail.com (G.M.M.).

Notes

The authors declare no competing financial interest.

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.*, in press; DOI:10.1002/minf.201200120.
(6) 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, DC, 2001; abstract no. 77.
(7) 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.
(8) Yongye, A. B.; Byler, K.; Santos, R.; Martínez-Mayorga, K.; Maggiora, G. M.; Medina-Franco, J. L. Consensus Models of Activity Landscapes with Multiple Chemical, Conformer, and Property Representations. *J. Chem. Inf. Model.* **2011**, *51*, 2427−2439.
(9) Willett, P. Searching Techniques for Databases of Two- and Three-Dimensional Structures. *J. Med. Chem.* **2005**, *48*, 4183−4199.
(10) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(11) Stumpfe, D.; Bajorath, J. Frequency of Occurrence and Potency Range Distribution of Activity Cliffs in Bioactive Compounds. *J. Chem. Inf. Model.* **2012**, *52*, 2348−2353.
(12) 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.
(13) Feller, W. *An Introduction to Probability Theory and Its Applications*,3rd ed.; John Wiley and Sons: 1968; Vol. *I*, pp 7−242.
(14) In our previous work,[5] a different sample space consisting of *unique* compound pairs was used. This is discussed further in the text following eq 6.
(15) Note that the random variables are given as upper case nonitalicized letters to distinguish them from crisp sets, which are designated by upper case italicized letters.
(16) Technically, the events are the inverse or preimages of two different partitionings of the real number line associated with the random variables S and A given by the threshold values s_T and a_T. Hence, $X^{<} \cup X^{\geq} = \mathbf{\Omega}$ and $X^{<} \cup X^{\geq} = \varnothing$ for $X = S,A$, that is $X^{<}$ and $X^{\geq}$ are complements of one another and partition the sample such that its cardinality is conserved, i.e. $|X^{<}| + |X^{\geq}| = |\mathbf{\Omega}|$ and the corresponding probabilities sum to unity.
(17) Zadeh, L. A. Fuzzy Sets. *Information Control* **1965**, *8*, 338−353.
(18) Zimmermann, H.-J. Fuzzy Set Theory. *WIRES Comput. Stat.* **2010**, *2*, 317−332.
(19) The tilde "∼" is again employed to emphasize that the sets corresponding to the three key regions of local SAS maps are fuzzy sets.
(20) Unlike the case in the theory of classical crisp sets, in fuzzy sets the operations of intersection "∩" and union "∪" can take on many forms such as the *algebraic product* employed in this work for set intersection. See e.g. Kaufmann, A. *Theory of Fuzzy Subsets*; Academic Press: New York, 1975; p 35.
(21) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities. *Nucleic Acids Res.* **2007**, *35*, D198−D201.
(22) 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.
(23) Peltason, L.; Bajorath, J. SAR Index: Quantifying the Nature of Structure-Activity Relationships. *J. Med. Chem.* **2007**, *50*, 5571−5578.