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).

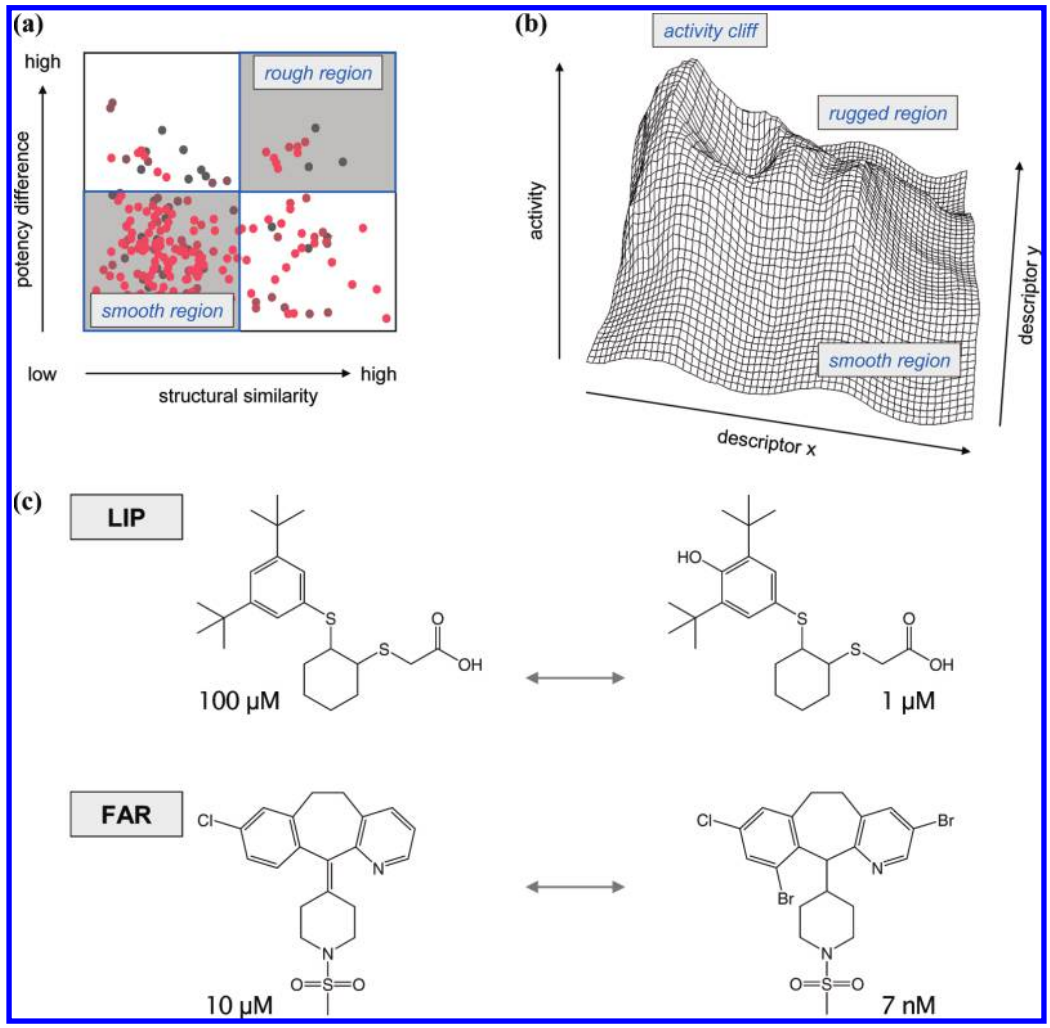

Figure 1. Schematic activity landscape representations and activity cliffs. (a) Similarity−potency plot. Pairwise structural similarity of active molecules is plotted against differences in logarithmic potency. Each data point represents a pairwise compound comparison and is colored according to the 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 nxn matrix of pairwise Euclidean distances δ_{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 δ_{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 δ_{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*max − *x*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:

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

Here, P_i denotes the negative decadic logarithm of the potency value of compound *i*, and δ_{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.

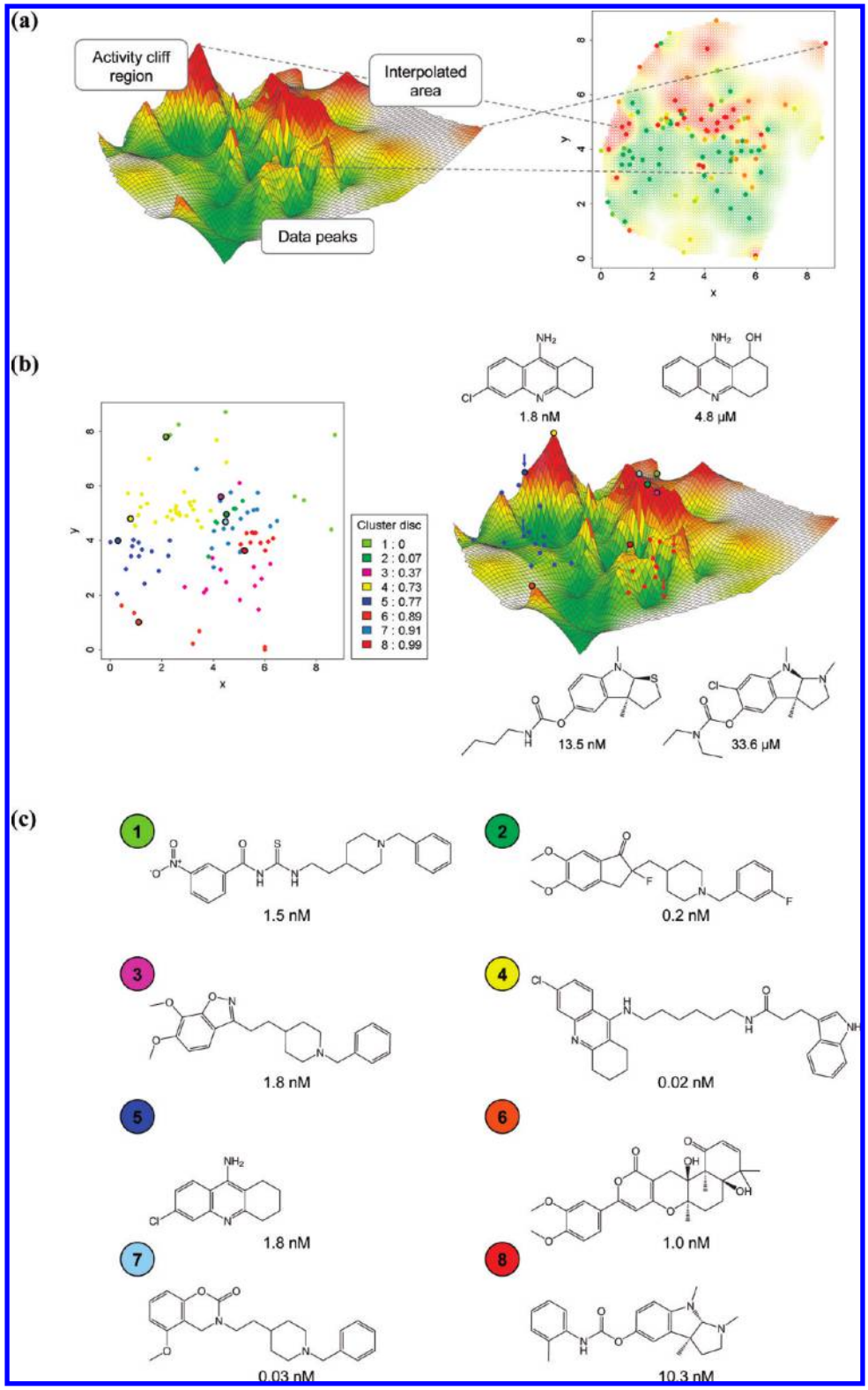

Figure 2. Interpretation of activity landscape representations. For the ACH data set and MACCS fingerprints, 2D and 3D activity landscape representations are shown. (a) Comparison of 2D and 3D landscape. The 3D landscape (left) contains distinct regions that are discussed in the text. These regions can 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.

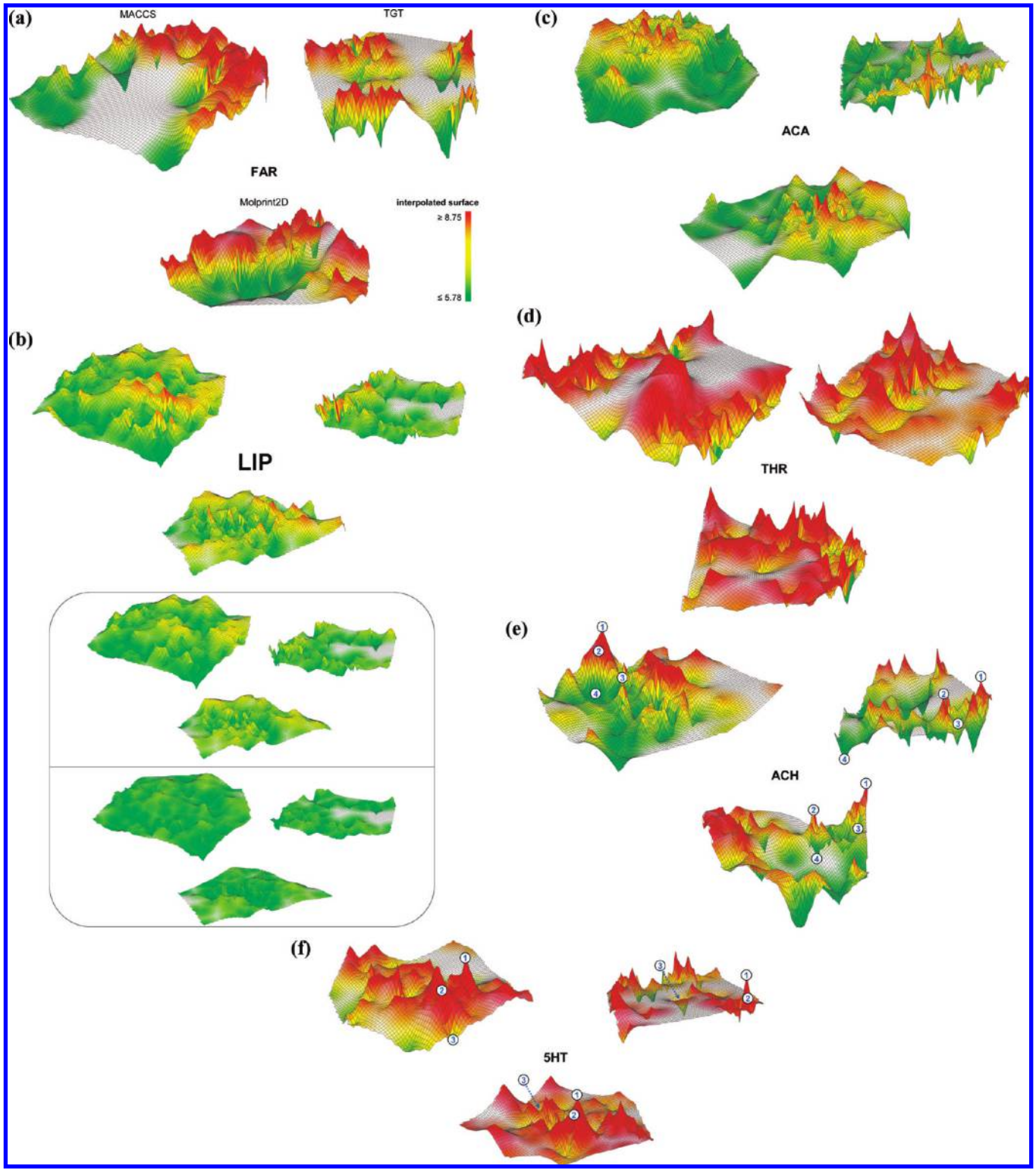

Figure 3. Activity landscapes. For the six compound data sets according to Table 1, activity landscapes were calculated on the basis of Euclidean fingerprint distances for three fingerprint representations, MACCS (top left), TGT (top right), and Molprint2D (bottom). The surface is colored according to 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

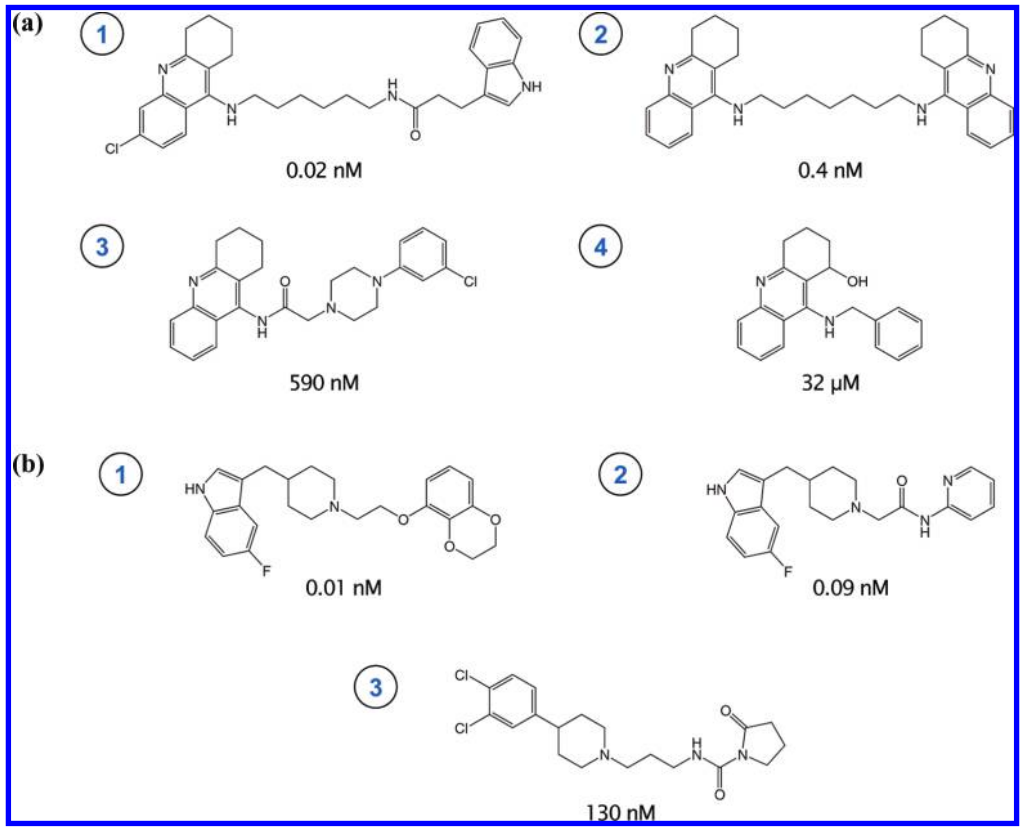

Figure 4. Exemplary compounds. For (a) ACH and (b) 5HT molecules are shown that are labeled in the activity landscapes in Figure 2e and Figure 2f, respectively. Depending on the chosen fingerprint representation, these molecules map to different regions of the landscapes and form, or do not form, activity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

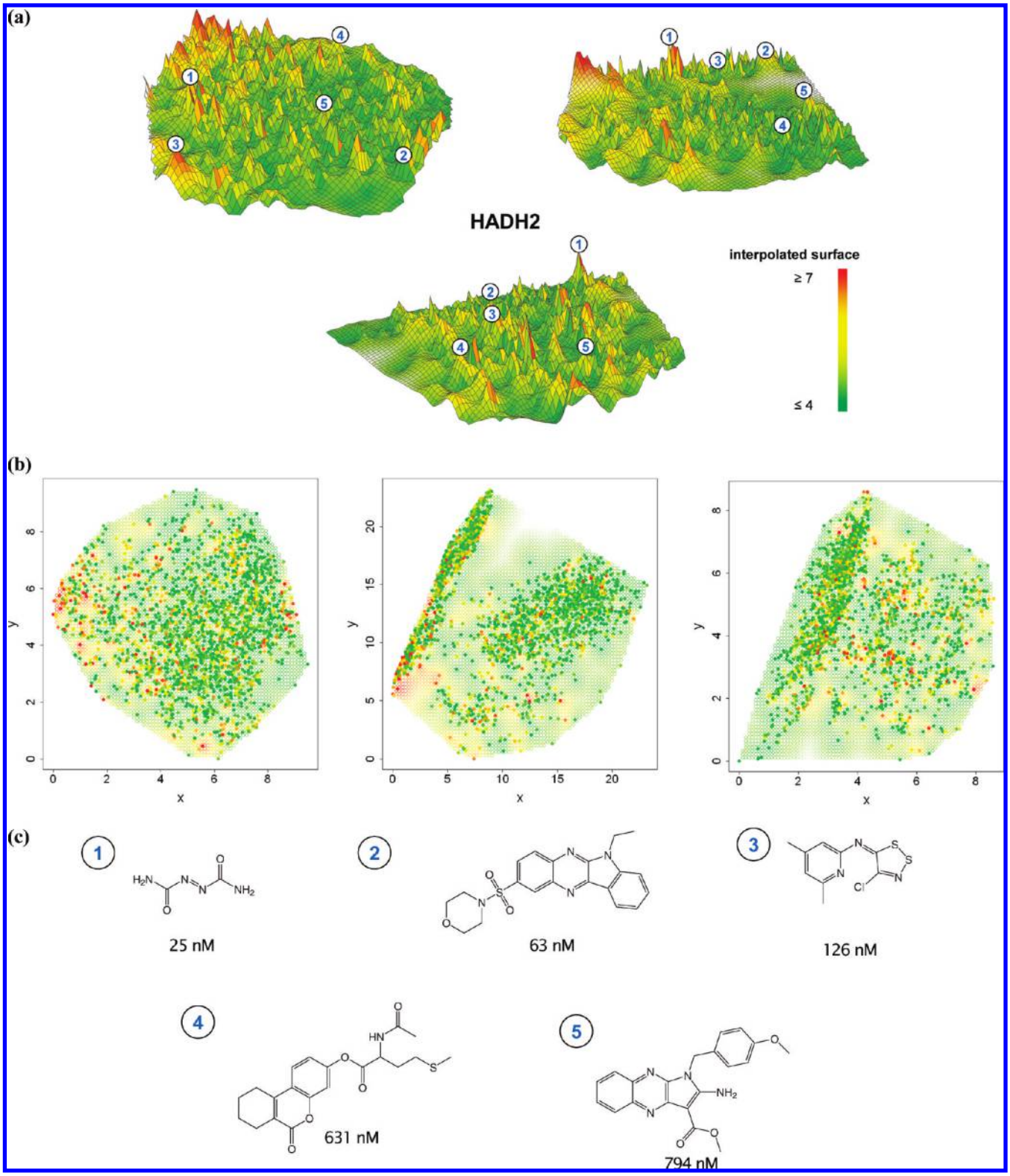

Figure 5. Activity landscape for HTS data. Activity landscape representations for a set of 2398 inhibitors of hydroxyacyl−CoA dehydrogenase II taken from a screening set are shown for three different fingerprint representations. (a) 3D landscape representations for MACCS (left), TGT (right), and Molprint2D (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