Journal of Medicinal Chemistry Perspective

J. Med. Chem. **2010,** *53,* 8209–8223 **8209**
DOI: 10.1021/jm100933w

Activity Landscape Representations for Structure−Activity Relationship Analysis

Anne Mai Wassermann, Mathias Wawer, 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 July 23, 2010

Introduction

The study of compound structure−activity relationships (SARs[a]) is one of the central themes in medicinal chemistry. SAR information is analyzed in different contexts, from screening and hit-to-lead to lead optimization projects. For the exploration of SARs, the concept of an activity landscape, which integrates molecular similarity and potency information, is of high relevance. The computational study of activity landscapes is still an evolving field. Activity landscape models are designed to rationalize SAR features of compound data sets and select key compounds for chemical exploration. The choice of molecular representations and the way molecular similarity is assessed are critically important factors for landscape generation and analysis. Graphical representation of SAR features is a major focal point of landscape modeling. Although complex activity landscapes are generally difficult to analyze, much progress has recently been made in extracting SAR information from various landscape views. This Perspective aims to provide an overview of the state-of-the-art in activity landscape analysis and a discussion of its potential for medicinal chemistry applications.

Understanding how structural modifications affect the biological activity of compounds or deriving a pharmacophore hypothesis from diverse active chemical entities present challenges that can be tackled using medicinal chemistry experience and intuition and/or computational tools. By no means is SAR analysis a priori dependent on computational methods. Rather, SAR analysis is often carried out on paper or whiteboards, by comparing molecular graphs of active compounds, consistent with the way chemists are traditionally trained. It has been pointed out that judgments of medicinal chemists are naturally *subjective* and often inconsistent.[1] This is of course not specific to medicinal chemistry but rather a consequence of how we as individuals subjectively access and evaluate data sets of any kind.[1] Likely inconsistencies in individual judgments about chemical and biological data might well be taken as an argument to promote the use of computational methods for SAR analysis. However, it would be rather careless to assume that computational analysis would per se be *objective*. In fact, computational objectivity does not exist. We typically apply models with underlying assumptions and inherent approximations that are often only useful within relatively narrow applicability domains and the results of which are generally difficult to evaluate.[2] In this context, it is often overlooked that we can not model phenomena whose physicochemical or biological foundations we do not understand. Of course, calculations that are carried out and reported should at least be reproducible (one would hope), but reproducibility does not mean objectivity.

There is, however, a rather simple factor that generally favors computational approaches to SAR analysis, and that is data set size. As long as one investigates one compound series at a time, knowledge of chemical graphs and activity data might be readily sufficient to deduce and predict SAR behavior. However, as molecular data sets grow in size, we quickly approach our limits to access and compare structures and associated biological properties such that computational data processing and analysis often become essential. Many compound data sets that have accumulated in pharmaceutical settings go far beyond the capacity of medicinal chemistry-centric SAR analysis and require the application of specialized computational tools for data handling and also modeling. Again, given the model-based nature of computational SAR analysis schemes, this does not make SAR analysis necessarily more objective (than individual assessments), but it makes it feasible.

Currently available computational approaches to SAR analysis are multifaceted and of rather different methodological complexity. A general distinction can be made between methodologies that primarily help to access and visualize SAR data obtained from screening[3,4] or chemical optimization[5] campaigns and those that ultimately predict biological activities. Among predictive methods, there are, for example, approaches to model linear and nonlinear[6,7] structure−activity relationships, in particular, those based on the classical QSAR paradigm,[7] pharmacophore techniques,[8] and various machine learning approaches.[9] Activity landscape methods, as introduced in the following, add to this methodological spectrum a strong focus on data-driven, descriptive, and large-scale SAR analysis schemes.[10]

The Activity Landscape Concept

The complexity and vastness of chemical space and the biological relevance of small subspaces or "islands" have been much discussed.[11,12] In computational medicinal chemistry, we do not have a unified space of the chemical universe available.

*To whom correspondence should be addressed. Phone: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

[a] Abbreviations: CAG, combinatorial analogue graph; NSG, network-like similarity graph; MMP, matched molecular pair; QSAR, quantitative structure−activity relationship; SAR, structure−activity relationship; SAS map, structure−activity similarity map; SPT, similarity-potency tree; 2D, two-dimensional; 3D, three-dimensional; SALI, structure−activity landscape index; SARI, structure−activity relationship index.

Published on Web 09/16/2010

Rather, computational chemical reference spaces must usually be generated with the aid of numerical descriptors of chemical structure and molecular properties,[9] which are in part rather abstract formulations. There are many arbitrary and subjective elements involved in chemical space design, and generally applicable computational space representations do not exist.[9] For activity prediction, the holy grail of computational space design is activity relevance. A chemical reference space is suitable for activity prediction if distance relationships between test compounds correlate with their biological properties. This means that compounds having similar activity should be close in such chemical reference spaces but distant from inactive compounds or other activity classes.

In the context of SAR analysis, activity landscape modeling is carried out for sets of specifically active compounds having different potency. Hence, in this case, requirements of chemical reference spaces differ from those utilized for activity prediction in that distance relationships in chemical space should predominantly reflect structural similarity.

In principle, however, derived chemical space is transformed into an activity landscape by adding an activity hypersurface to it that accounts for differences in compound potency. The underlying concept is intrinsically simple. Compound positions in chemical space are "decorated" with potency information. Structurally similar compounds map close to each other, structurally distinct compounds are far apart, and all compound potency differences are reflected by the activity hypersurface. Of course, in high-dimensional space representations, we are unable to directly access and interpret structure−potency relationships and hence a critically important aspect of activity landscape design is to generate interpretable 2D or 3D representations of landscapes for given data sets.

In general terms, we can define an activity landscape as any representation that integrates the analyses of the structural similarity of and potency differences between compounds sharing the same biological activity. This definition covers rather different types of landscape representations, as discussed in the following.

Characteristic features of activity landscape analysis include its data-driven (what do available compound activity data tell us?) and descriptive (rather than predictive) nature.[10] Informative activity landscape views should provide an intuitive access to SAR information that might otherwise be difficult to obtain, as illustrated in the following. Importantly, activity landscape analysis will not replace chemical and knowledge-based interpretation of SAR features but provide an advanced basis for it.

Origins of Activity Landscape Modeling

The analysis of activity landscapes is conceptually related to the study of general chemical "neighborhood behavior",[13] i.e. the way calculated molecular similarity relates to the biological activity of test compounds in quantitative terms. Much of the early and pioneering work on activity landscapes was carried out by Gerald M. (Gerry) Maggiora and his colleagues at what was then Upjohn, and later on Pharmacia Corporation, in Kalamazoo, Michigan. In 2001, Shanmugasundaram and Maggiora presented structure−activity similarity (SAS) maps, a prototypic 2D activity landscape representation.[14] A schematic SAS map is shown in Figure 1. For a given set of compounds, SAS maps compare structural similarity and "activity similarity" on the basis of systematic pairwise

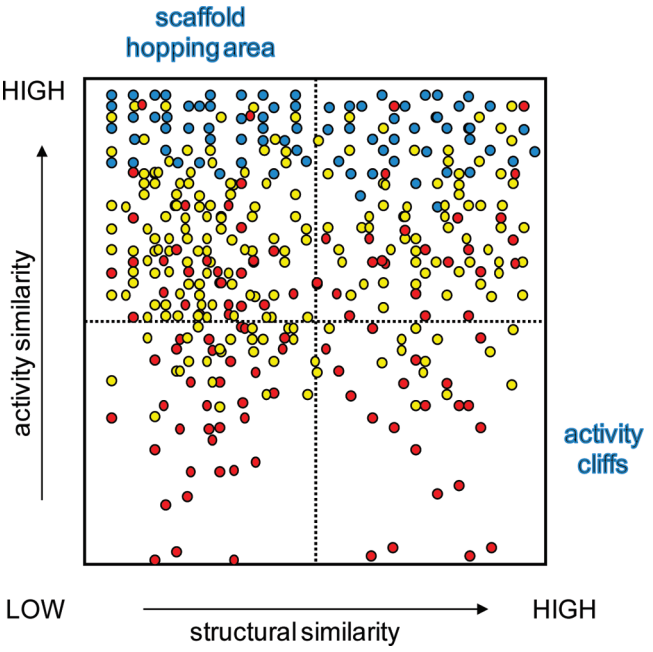

Figure 1. SAS map. For all compound pairs within a data set, structural similarity is plotted against activity similarity. Each data point (colored dot) corresponds to a pairwise compound comparison. Four regions of different SAR character can essentially be distinguished in an SAS map. The upper-left section contains structurally diverse compounds with similar activity and corresponds to a scaffold hopping area, whereas the lower-right section contains activity cliff-forming compound pairs, i.e. structurally similar compounds with a significant difference in potency. The color code of the data points adds a further level of information by indicating whether the more active compound in a pair is highly (blue), intermediately (yellow), or only weakly (red) potent. The example shown represents a hypothetical compound data set.

compound comparisons. For this purpose, activity similarity is defined for two compounds i and j as follows:

$$\mathrm{sim}_{\mathrm{act}}(i,j) = 1 - \frac{\left|P_i - P_j\right|}{P_{\max} - P_{\min}} \qquad (1)$$

Here P_i gives the potency of compound i (for example, as pK_i or pIC_{50}) and $P_{\max} - P_{\min}$ the difference between the maximum and minimum potency observed in the compound set. For any compound pair, a normalized potency difference is obtained and each data point in the SAS map represents a pairwise compound comparison. Data points are color-coded according to the potency value of the more active compound of each pair. Compound similarity can be calculated in different ways, as further discussed below. This type of similarity−potency representation can be modified in many ways. For example, compound potency differences can be plotted instead of activity similarity and data points can be color-coded according to the sum of compound potency values, taking into account the potency range within the data sets.[15]

A key feature of an SAS map is that it can be understood to consist of four sections capturing different principal features of an activity landscape, as indicated in Figure 1. The upper-left section is populated by compound pairs with high activity similarity and low structural similarity. Thus, this region corresponds to a "scaffold hopping"[16] area where diverse structures have similarly high or low activity. Here, only

compounds with high potency are of significant interest. The upper-right section of the map contains compound pairs with high structural and high activity similarity. These compounds might represent, for example, series of analogues with comparable potency. Less interesting are compound pairs falling into the lower-left section having low structural and low activity similarity. By contrast, compound pairs in the lower-right region have high structural similarity but low activity similarity. Hence, these are compounds, often series of analogues, where small structural modifications lead to significant changes in potency. In activity landscape terminology, high structural−low activity similarity compound pairs are referred to as "activity cliffs", a designation that will become rather intuitive when 3D landscape views are considered. Such activity cliff regions are most difficult to navigate for many (but not all) computational methods, which makes them interesting from more than one point of view. In the relevant literature, the term "activity cliff" can be traced back to a book chapter published in 1991 by Michael S. Lajiness,[17] then a colleague of Gerry Maggiora at Upjohn-Pharmacia.

Idealized 3D Activity Landscapes and Activity Cliffs

Maggiora and colleagues also realized early on the attractiveness of considering activity landscapes as topographical maps reminiscent of actual geographical landscapes.[18] These maps correspond to theoretical 3D landscapes with idealized topology, as shown in Figure 2. These idealized 3D activity landscapes can be rationalized as a 2D projection of the chemical space representation ($x-y$ plane) with compound potency added as a third dimension. In a landscape model, the hypothetical potency value distribution is represented as a contiguous surface (corresponding to a biological hypersurface in chemical space, as discussed above). Importantly, these landscape models provide an intuitive access to fundamental SAR characteristics. The model shown in Figure 2a contains smooth and gently sloped regions, whereas the landscape in Figure 2b contains rugged areas that represent activity cliffs. On the basis of idealized 3D activity landscapes, it can be well appreciated that activity cliff areas represent the most prominent features of an activity landscape. Moreover, we can deduce principal SAR characteristics from idealized landscape topology. In gently sloped rolling hill-like regions, gradual structural changes are accompanied by only small to moderate changes in compound potency, and increasingly diverse structures fall within the same potency range. Hence, these areas correspond to regions of "SAR continuity".[15,19] In principle, focusing on subsets of compounds populating such regions makes scaffold hopping via virtual screening methods[9] a promising approach. Furthermore, linear modeling of SARs and activity predictions on the basis of linear models, strongly depend on the presence of SAR continuity among active compounds. In contrast to gently sloped areas of activity landscapes, activity cliff regions correspond to "SAR discontinuity". Here, small changes in compound structure are accompanied by large-magnitude changes in potency. Accordingly, compound subsets populating activity cliff regions severely limit the applicability of standard QSAR models.[20] The landscape shown in Figure 2c is a so-called "variable activity landscape".[19] Such variable activity landscapes correspond to the presence of "SAR heterogeneity", i.e. the combination, or coexistence, of continuous and discontinuous SAR components.

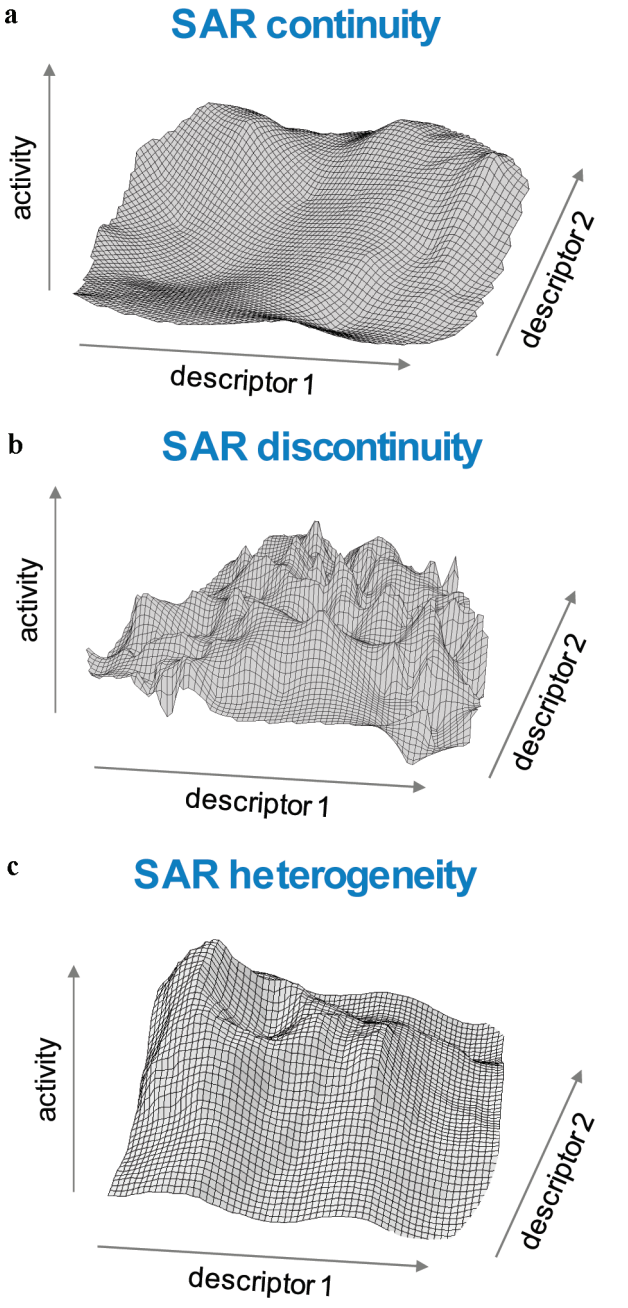

Figure 2. Hypothetical 3D activity landscapes. A 3D activity landscape adds potency as a third dimension to a set of compounds in a 2D projection of chemical reference space. Here potency distributions are hypothetical and three idealized hypersurfaces representing different SAR characters are shown: (a) SAR continuity, (b) discontinuity, and (c) heterogeneity. In the 2D projection of original chemical space, distances between compounds account for their dissimilarity. From potency values of individual compounds recorded on the vertical axis, a coherent surface must be generated through interpolation.

Rationalizing SAR Information Content

The question of what represents SAR information is more complicated to answer than it might appear at first glance. Clearly, for a practicing medicinal chemist, useful SAR information should reveal trends how to best make the next molecule(s). Thus, the primary focus is on understanding how structural modifications change compound potency in a defined and, ultimately, predictable manner. However, this essentially requires the presence of appreciable SAR continuity

and contrasts with SAR information content from an information-theoretic point of view. Here, information entropy (SAR information content) would be highest if potency values were randomly distributed over a chemical reference space—and thus be essentially unpredictable. Similarly, in activity landscapes, activity cliffs that represent the extreme form of SAR discontinuity are generally thought to represent regions of highest SAR information content[14] because small structural changes cause large differences in the biological response. Compounds forming activity cliffs reveal substitution sites critical for compound potency but not necessarily trends that help to understand how to further improve an active compound. Accordingly, there is an apparent discrepancy between high SAR information content associated with activity cliffs or "information-rich" potency value distributions on one side and chemical interpretability of SAR features (and SAR modeling) on the other. Accordingly, for the purpose of activity landscape analysis, we need to distinguish between chemical accessible and interpretable SAR information and different levels of SAR information content associated with activity cliff regions. This also implies that approaches to bridge between SAR continuity and discontinuity should be useful to sample SAR information from rather different points of view, as further discussed later on.

The Similarity Caveat

Activity landscape design only requires compound similarity relationships and potency values as input. While potency values are obtained from experiment, molecular similarity needs to be calculated. Although potency data are frequently affected by measurement errors, the assessment of molecular similarity presents the most critical variable for landscape modeling and a major source of potential inconsistencies in describing and comparing activity landscapes. In computational medicinal chemistry, compound similarity is evaluated in different ways. For example, in the context of pharmacophore or QSAR analysis, local similarity measures are typically applied by focusing on arrangements of substructures or functional groups in molecules that are activity determinants. By contrast, methods that conceptually rely on the "similarity property principle" (i.e., similar molecules should have similar biological properties)[21] employ whole-molecule similarity measures. This is usually also the case in activity landscape modeling (except if only series of closely related analogues are studied). Importantly, the assessment of whole-molecule similarity is significantly influenced by the molecular representations that are chosen (usually more so than by alternative similarity or distance metrics). Alternative molecular representations such as, for example, different fingerprints or combinations of different numerical molecular property descriptors, correspond to different chemical reference spaces. Similarity relationships are typically space-dependent and might vary considerably by moving from one reference space into another. Such variations can significantly change the topology of activity landscapes and their information content. For example, high-resolution descriptors that overemphasize small chemical changes might flatten, or even eliminate, activity cliffs that are found in chemically more realistic reference spaces. In addition, compounds that are similar in one feature space might be considered dissimilar when different properties are evaluated. Consequently, activity cliffs have been studied in alternative chemical space representations in order to identify "consensus activity cliffs" that are consistently formed in different reference spaces.[22] From a medicinal chemistry point of view, such cliffs would certainly be regarded as the most reliable ones. However, one should also consider that the application of consensus similarity or consensus scoring methods might often lead to eliminating data from further consideration that are (perhaps inappropriately) disfavored by an individual method. Nevertheless, the search for activity cliffs that are formed in landscapes resulting from different molecular representations and/or similarity methods is an attractive approach.

Importantly, as further illustrated below, we need to take into account that alternative molecular representations might profoundly change similarity relationships and affect the topology and interpretability of activity landscape models, much more so than limited experimental errors in potency measurements.

Numerical SAR Analysis Functions

Large-scale analysis of SAR features contained in compound data sets has been facilitated through the introduction of numerical SAR analysis functions including the SAR index (SARI)[15,23] and the structure−activity landscape index (SALI)[24] reported by Guha and van Drie (having its origins also in the former Upjohn-Pharmacia environment[24]). These analysis functions systematically evaluate, and score, pairwise similarity and compound potency relationships in compound data sets and thus directly access and mirror activity landscape features.

SARI is composed of two separately calculated scores, the continuity score and the discontinuity score. The raw continuity score is calculated as the potency weighted arithmetic mean of pairwise compound dissimilarity within a set A. The continuity score strongly weights structurally diverse compounds having high potency and small differences in potency. Thus, it accounts for gently sloped regions of an activity landscape:

$$\mathrm{cont}_{\mathrm{raw}}(A) = \underset{\{(i,j)\in A|i\neq j\}}{\text{weighted mean}}\left(\frac{1}{1+\mathrm{sim}(i,j)}\right) = \frac{\sum_{\{(i,j)\in A|i\neq j\}}\left(\mathrm{weight}(i,j)\cdot\frac{1}{(1+\mathrm{sim}(i,j))}\right)}{\sum_{\{(i,j)\in A|i\neq j\}}\mathrm{weight}(i,j)}$$

$$\mathrm{weight}(i,j) = \frac{P_i\cdot P_j}{1+|P_i-P_j|} \quad (2)$$

where P stands for potency and $\mathrm{sim}(i,j)$ for the similarity of compounds i and j (that is usually calculated as Tanimoto similarity of fingerprint representations).

The raw discontinuity score is calculated as the average pairwise potency difference between compounds multiplied by pairwise similarity:

$$\mathrm{disc}_{\mathrm{raw}}(A) = \underset{\{(i,j)\in A|\mathrm{sim}(i,j)>T,\,|P_i-P_j|>1\}}{\text{mean}}(|P_i-P_j|\cdot\mathrm{sim}(i,j)) \quad (3)$$

The discontinuity score emphasizes structurally similar compounds with large potency differences and hence accounts for rugged regions of an activity landscapes and activity cliffs. Because the discontinuity score is designed to monitor the

presence of activity cliffs, only pairs of compounds with at least 1 order of magnitude difference in potency and a similarity exceeding a predefined threshold T are considered.

The raw scores are converted into Z-scores using the score distribution of a reference panel of compound sets.[15] Then Z-scores are mapped onto the value range [0,1] by calculating the cumulative probability for each Z-score under the assumption of a normal distribution. The resulting (normalized) continuity and discontinuity scores are combined to yield the SARI value:

$$\mathrm{SARI}(A) = \frac{1}{2}(\mathrm{cont}_{\mathrm{norm}}(A) + (1 - \mathrm{disc}_{\mathrm{norm}}(A))) \quad (4)$$

The final SARI score balances SAR continuity and discontinuity contributions. In its original implementation, it was applied to quantify the global SAR character of compound data sets and classify global SARs into three categories; i.e. continuous (high SARI scores), discontinuous (low scores), or heterogeneous (intermediate scores around 0.5) SARs. Extensive profiling of different sets of active compounds revealed that the majority of global SARs are heterogeneous in nature,[15] consistent with the presence of variable activity landscapes, as illustrated in Figure 2c. Such SAR heterogeneity can arise from the mutual coexistence of continuous and discontinuous SARs in different compound subsets or from the presence of continuity in the vicinity of an activity cliff (for example, when structural variations in active compounds are permitted as long as one or more strong binding constraints are met).

Importantly, SARI discontinuity scoring can also be carried out on a per-compound basis, yielding a "local", rather than "global", score. Compound discontinuity score calculation is carried out by comparing a compound to all other molecules that are more similar to it than the predefined threshold T. Scores are normalized by using the individual scores of all compounds in the data set as a reference for Z-score calculations (instead of an external reference panel).[23] The compound SARI discontinuity score then accounts for contributions of individual compounds to the introduction of local SAR discontinuity.

The SALI scoring scheme is designed to quantify activity cliffs and calculated as follows:

$$\mathrm{SALI}(i,j) = \frac{P_i - P_j}{1 - \mathrm{sim}(i,j)} \quad (5)$$

Thus, SALI essentially corresponds to the SARI discontinuity score. Differences between this and the SARI discontinuity score include that SALI is a pairwise score with infinite value range that emphasizes large potency differences between similar compounds, whereas the SARI discontinuity score takes average potency differences of all pairs of similar compounds into account and is normalized. Because of their local nature, average SALI scores can not be utilized as a measure of SAR heterogeneity. For this purpose, a global scoring scheme must be applied.

Activity Cliff-centric Landscape Views

The SALI formalism can be elegantly applied to produce activity cliff-centric representations of activity landscapes. In SALI graph representations, nodes represent compounds and edges activity cliffs, i.e. two compounds are connected if their SALI score exceeds a predefined threshold value (i.e., greater than 50%, 60%, 70%, ... of all scores).[24] Then, activity cliffs of increasing magnitude can be identified. Edges are directed according to increasing compound potency. Figure 3 shows a prototypic SALI map. Calculation of graph representations at increasing SALI score threshold levels identifies series of pairwise connected activity cliffs that might often represent compound optimization pathways, an attractive application for medicinal chemistry. The SALI score threshold is critical for a meaningful assessment of activity cliffs. If it is set to low values, a data set is considered to contain a continuum of activity cliffs.

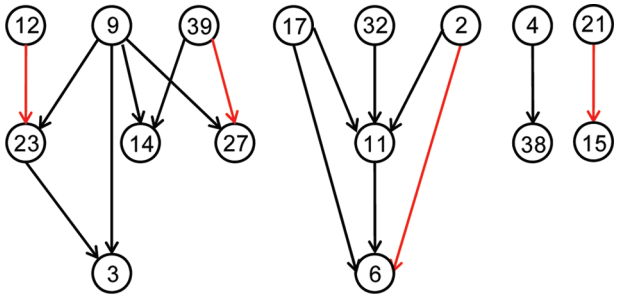

Figure 3. Schematic SALI graph representation. Compounds are displayed as nodes and labeled with identifiers. Pairs of compounds are connected by an edge if their SALI value exceeds a user-specified threshold. Edges are directed according to increasing compound potency. With an increasing SALI score level threshold, only a few edges remain (red) that connect compound pairs, forming the most significant activity cliffs in the data set.

Another elegant application of the SALI approach is to analyze how many edges in a SALI graph (i.e., activity cliffs) are correctly accounted for by different SAR models.[25] This type of analysis makes it possible to prioritize alternative computational models for application to compound data sets containing different SAR information. For this purpose, a SALI curve is generated that reports the fraction of correctly predicted pairwise compound potency relationships (directed edges) as a function of the SALI map score threshold value. The more directed edges are predicted by a given model at increasing activity cliff stringency, the better it is suited for handling the data set. Like any activity landscape analysis, this assessment is also influenced by the chosen molecular representations and similarity methods.

Guha recently also proposed a modification of the SALI formalism by taking compound dose−response behavior into account instead of single-point potency measurements.[26] This can be accomplished by replacing potency differences in the numerator of the SALI formula with the Euclidian distance between Hill equation parameters of dose−response curves of the compared compounds. The modification is thought to further refine the description of activity cliffs of moderate magnitude.[26]

Activity Cliff Distribution

As discussed above, the formation of activity cliffs in compound data sets can well be considered a continuum, but activity landscape analysis primarily focuses on identifying the most prominent cliffs. Systematic activity landscape analyses have revealed that most, if not all, sets of active compounds, also including screening data sets, contain activity cliffs of moderate to large magnitude.[23,27] The formation of activity cliffs in series of structurally similar compounds generally results from different R-groups (substitutions). Hence, one might ask the question whether chemical substitutions exist

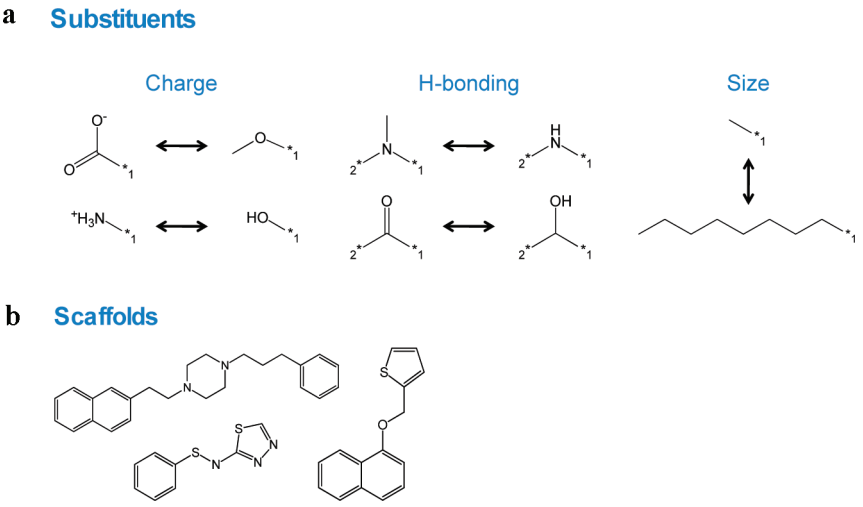

Figure 4. Activity-cliff inducing chemical substitutions and molecular scaffolds. (a) R-group replacements that frequently induce activity cliffs in different structural environments and across diverse biological targets are annotated with molecular property changes they correspond to. (b) Shown are three representative scaffolds that are contained in different compound pairs forming activity cliffs for multiple targets.

that display a general tendency to introduce activity cliffs. Considering the general requirements of productive receptor−ligand interactions, it would probably be expected that substitutions that, for example, introduce opposite charges or hydrophobic groups of different size might be frequent activity cliff inducers. Indeed, a systematic analysis of activity cliff-inducing substitutions on the basis of matched molecular pairs[28] identified ~200 R-group replacements having a general tendency to form activity cliffs in different compound classes active against different targets.[29] Figure 4a shows representative examples of such substitutions. It is important to note that activity cliff-forming substitutions are derived here exclusively on the basis of compound structure and potency comparisons, which does not take structural insights into receptor−ligand interaction into account. Of course, activity cliffs could also be rationalized on the basis of complex crystal structures of analogue series, although the structural approach would essentially be target-centric (and there is much less information available). Nevertheless, on a case-by-case basis, structural interaction analysis can rationalize activity cliff formation at the atomic level of detail.

In light of the strong focus on substitution patterns for activity cliff formation, a perhaps much less obvious question has been whether molecular frameworks (core structures) might also exist that frequently introduce activity cliffs. To address this question, active compounds representing unique molecular scaffolds have been systematically analyzed for their ability to form activity cliffs against different targets.[30] In this study, more than 100 scaffolds of varying chemical complexity have been identified that form significant cliffs across multiple (related or unrelated) targets. Representative examples are shown in Figure 4b. Taken together, these findings suggest that much can still be learned about the structural origins of activity cliff formation. Such insights would be expected to aid in the selection of compounds for chemical exploration (e.g., by identifying compounds representing preferred activity cliff scaffolds) and the design of optimization strategies (e.g., by evaluating substitutions having high cliff probability).

Global versus Local SARs

A landscape view that conceptually differs from SALI maps is provided by network-like similarity graphs (NSGs).[23] Here, the focus is on exploring relationships between the global SAR character of a compound data set and local SAR features. An exemplary NSG is shown in Figure 5a. Nodes represent compounds and are color-coded according to their potency values and edges represent similarity relationships (an edge is drawn between two nodes if their calculated pairwise 2D similarity exceeds a predefined threshold value). The size of nodes is scaled according to compound discontinuity scores. In addition to calculating pairwise similarity relationships, hierarchical clustering of the data set is carried out and the resulting clusters are highlighted. For each cluster, a cluster discontinuity score is also calculated. The relative arrangement of clusters to each other has no chemical meaning and is determined by a graphical layout algorithm. NSGs also provide an intuitive access to activity cliffs. Large red and green nodes connected by edges are activity cliff markers and indicate the most significant cliffs contained in a data set. Moreover, NSGs identify different local SAR environments. For example, the squalene synthase inhibitor set in Figure 5a is globally heterogeneous on the basis of SARI profiling, and the NSG reveals the presence of both strongly continuous and discontinuous compound clusters. By contrast, the thrombin inhibitor set shown in Figure 5b is characterized by significant global SAR discontinuity and, accordingly, its NSG shows the presence of many large red and green nodes that dominate the activity landscape. Hence, in addition to activity cliffs, NSGs provide an immediate access to different local SARs in compound data sets. Furthermore, NSGs can also be used to search for SAR information in raw screening data.[27] An example is shown in Figure 5c. For the analysis of screening data, confirmatory screens are preferred because of their reduced error rates. Hit sets from screening campaigns typically consist of many predominantly weakly active and often structurally diverse compounds, which corresponds to the presence of SAR continuity on a global scale. However, a lesson learned from profiling many screening sets is that the activity landscapes of essentially all of these data sets contain regions (compound subsets) of SAR discontinuity,[27] as illustrated in Figure 5c. These regions often provide focal points for hit selection. Although NSG analysis is readily applicable to screening data, and usually informative in these cases, conclusions about local SARs are of course only meaningful

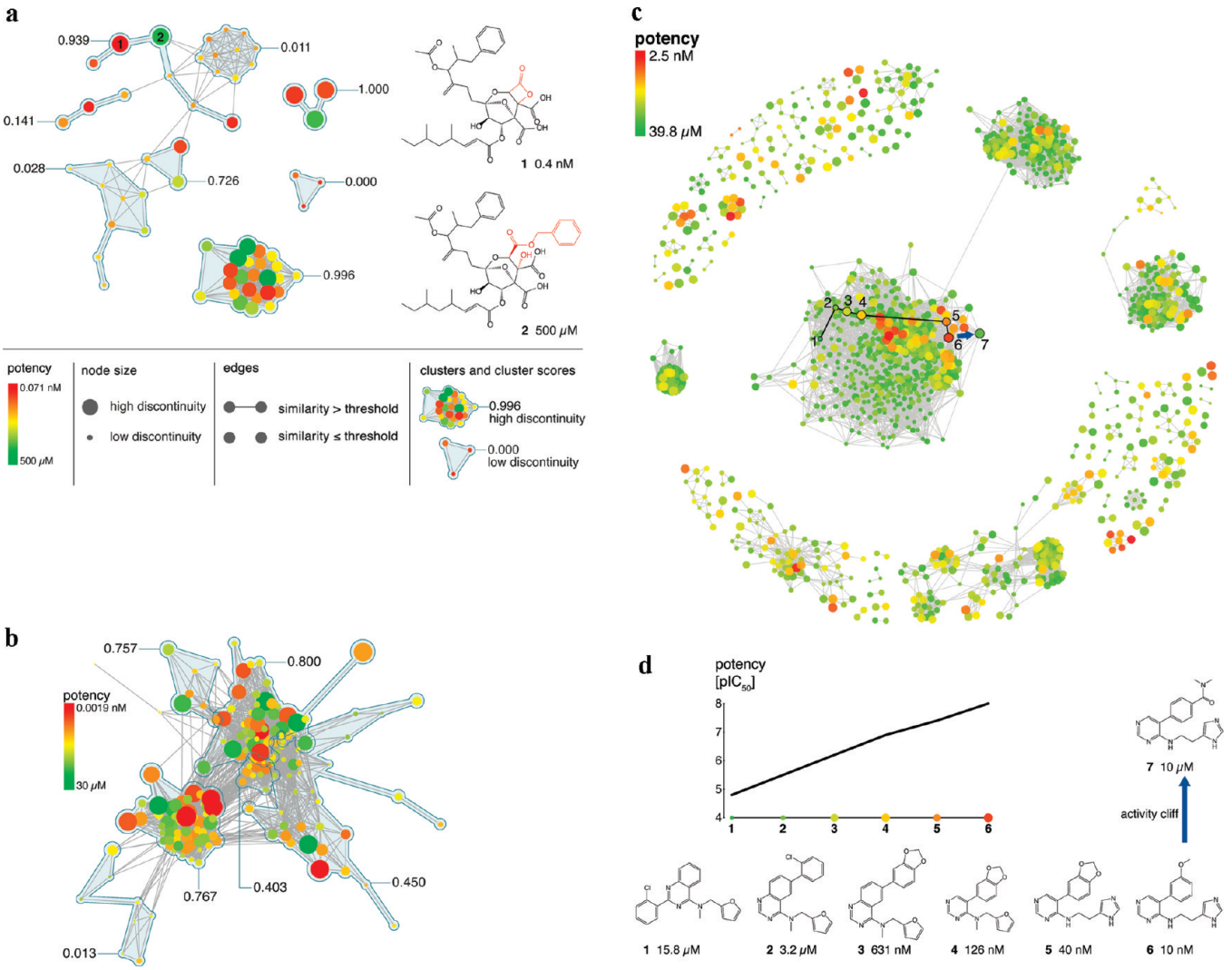

Figure 5. Network-like similarity graphs. (a) An NSG is shown for a set of 71 squalene synthase inhibitors. Nodes represent individual compounds and node colors reflect compound potency, as indicated by the color bar in the lower left corner. Edges indicate compound similarity relationships computed from 2D fingerprint representations. Node size is scaled according to compound discontinuity scores. Clusters of compounds are highlighted and annotated with cluster discontinuity scores. On the right, the two molecules are shown that form an activity cliff. The relative orientation of and distances between clusters are determined by a graphical layout algorithm. (b) NSG for a set of 172 thrombin inhibitors. Node and cluster annotations are analogous to (a). Parts (a) and (b) have been adapted from ref 23. (c) NSG for a set of 1379 cytochrome P450 isoform 2C19 screening hits (inhibitors). A compound pathway is highlighted. (d) Details of this pathway are shown in a similarity−potency diagram. The pathway is leading from a continuous local SAR to a discontinuous region. Compound **6** forms an activity cliff with compound **7** that is not part of the pathway.

if test compounds share the same specific activity. This would not be the case, for example, if it is not clear whether active compounds are receptor agonists or antagonists and when data sets contain mixtures of such compounds.

In Figure 5c, a compound path is highlighted that leads from a region of local SAR continuity to an activity cliff in a discontinuous region. This so-called "SAR pathway"[27] is represented in detail in Figure 5d. SAR pathways are based on a predefined SAR model and can be systematically computed for NSGs and ranked on the basis of their fit to the SAR model. This model prioritizes pathways that span a large potency interval between start and end compound and consist of as many as possible pairwise similar compounds following an ideally linear potency gradient with small potency increases between subsequent compounds. As such, the pathway model is designed to reflect SAR continuity. However, pathways can be identified that connect regions of SAR continuity to activity cliffs (by definition only a potent cliff marker can be the end point of the pathway), as shown in Figure 5d. The corresponding sequences of pairwise similar compounds may or may not represent compound optimization paths, but they provide a "chemical link" between SAR continuity and discontinuity. Because SAR pathways are intrinsically continuous in nature, they often provide interpretable information.

From Activity to Selectivity Cliffs

The NSG framework can also be utilized to study multitarget SARs resulting from compounds with activity against two or more targets. In this case, target selectivity of compounds results from different potencies against individual targets. In addition to the potency-based NSGs discussed above, selectivity-based NSGs can also be generated by using potency ratios (logarithmic potency differences) for two targets instead of individual compound potency values.[31] Figure 6a shows two potency-based NSGs for compounds

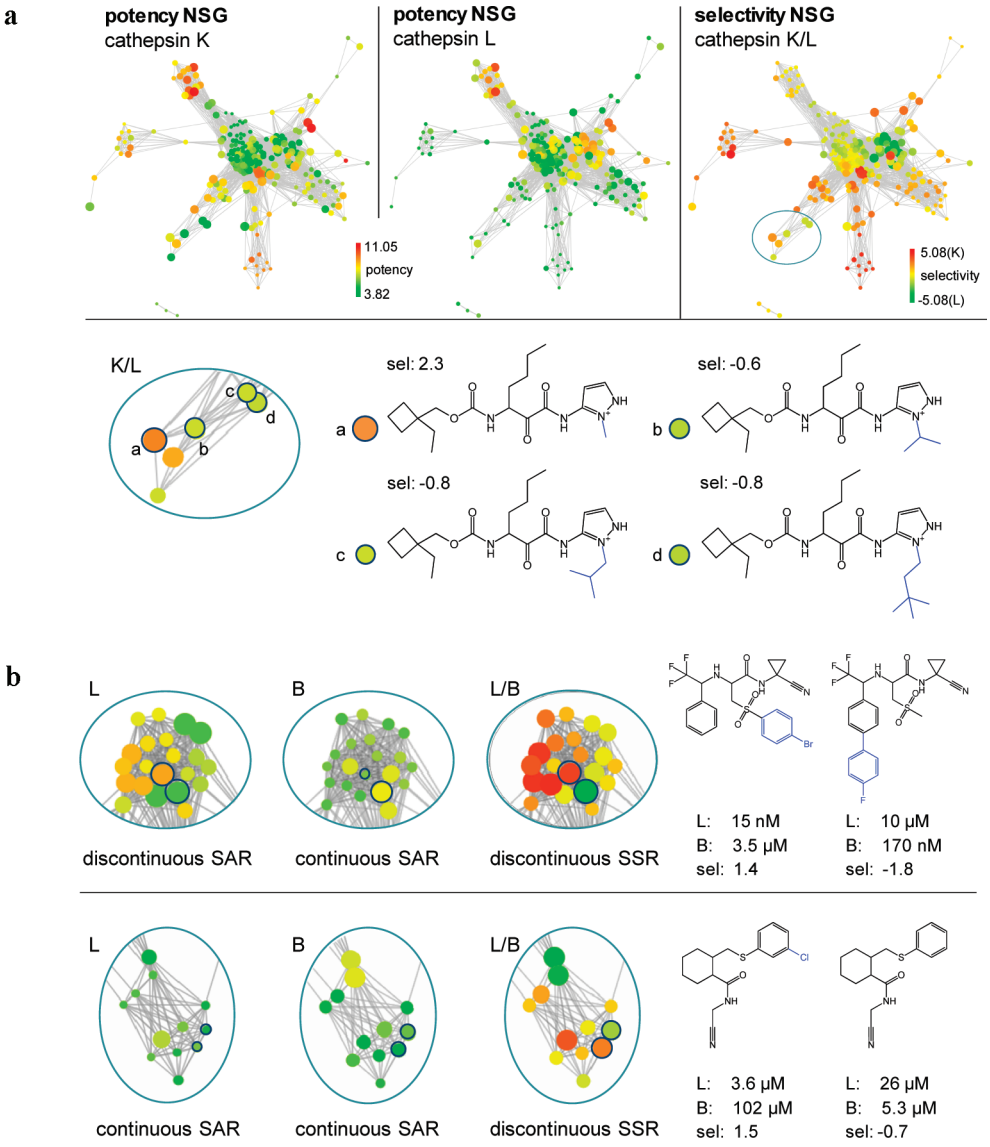

Figure 6. Potency and selectivity NSGs. (a) For a set of 234 inhibitors active against cathepsins K and L, the two potency NSGs for the individual targets and the selectivity NSG for the target pair are shown. By focusing on compounds in the vicinity of a selectivity cliff, selectivity determinants can be identified. Potency is reported as pK_i/pIC_{50} values and selectivity (sel) as the logarithmic potency differences for the two targets. (b) Corresponding local SAR and SSR environments from potency and selectivity NSGs for a set of 159 ligands active against cathepsins B and L are compared. Selectivity cliff markers that also mark, or do not mark, activity cliffs, are displayed and annotated with their potency and resulting selectivity for the two targets.

that inhibit both cathepsin K and L thiol proteases and the corresponding selectivity-based NSG. This representation makes it possible to study structure−selectivity relationships (SSRs).[31] The topology of the potency- and selectivity-based NSGs is the same because it is only determined by compound similarity relationships. In selectivity-based NSGs, compounds are color-coded according to the (logarithmic) selectivity range calculated for the data set. Here, combinations of large red and green nodes now represent "selectivity cliffs", i.e. compounds having markedly different selectivity against the two targets. As illustrated in Figure 6a, local SSR environments can be analyzed in selectivity-based NSGs (in analogy to local SAR environments). Structurally similar compounds in the vicinity of selectivity cliffs often reveal selectivity determinants, i.e. substitutions that significantly change relative potencies against the two targets.[31] Another attractive aspect of comparing potency- and selectivity-based NSGs is the exploration of relationships between activity and selectivity cliffs. Figure 6b shows corresponding local SAR and SSR environments from potency and selectivity NSGs of inhibitors of cathepsin B and L.[31] The example at the top in Figure 6b shows two compounds that participate in the formation of a discontinuous local SAR in the cathepsin L NSG where they form a moderately sized activity cliff. By contrast, the same compounds represent a continuous SAR in the corresponding region of the cathepsin B NSG. Moreover, the resulting local SSR environment in the selectivity NSG is characterized by strong discontinuity, and the two compounds form a steep selectivity cliff. Thus, in this case, activity cliff markers for cathepsin L, but not B, also form a selectivity cliff. The example at the bottom in Figure 6b shows a different relationship. In this case, two inhibitors of cathepsin L and B map to continuous local SAR regions in both potency NSGs, but the corresponding local SSR environment is discontinuous and the two compounds form a moderately sized selectivity cliff. Hence, depending on the potency ratios of

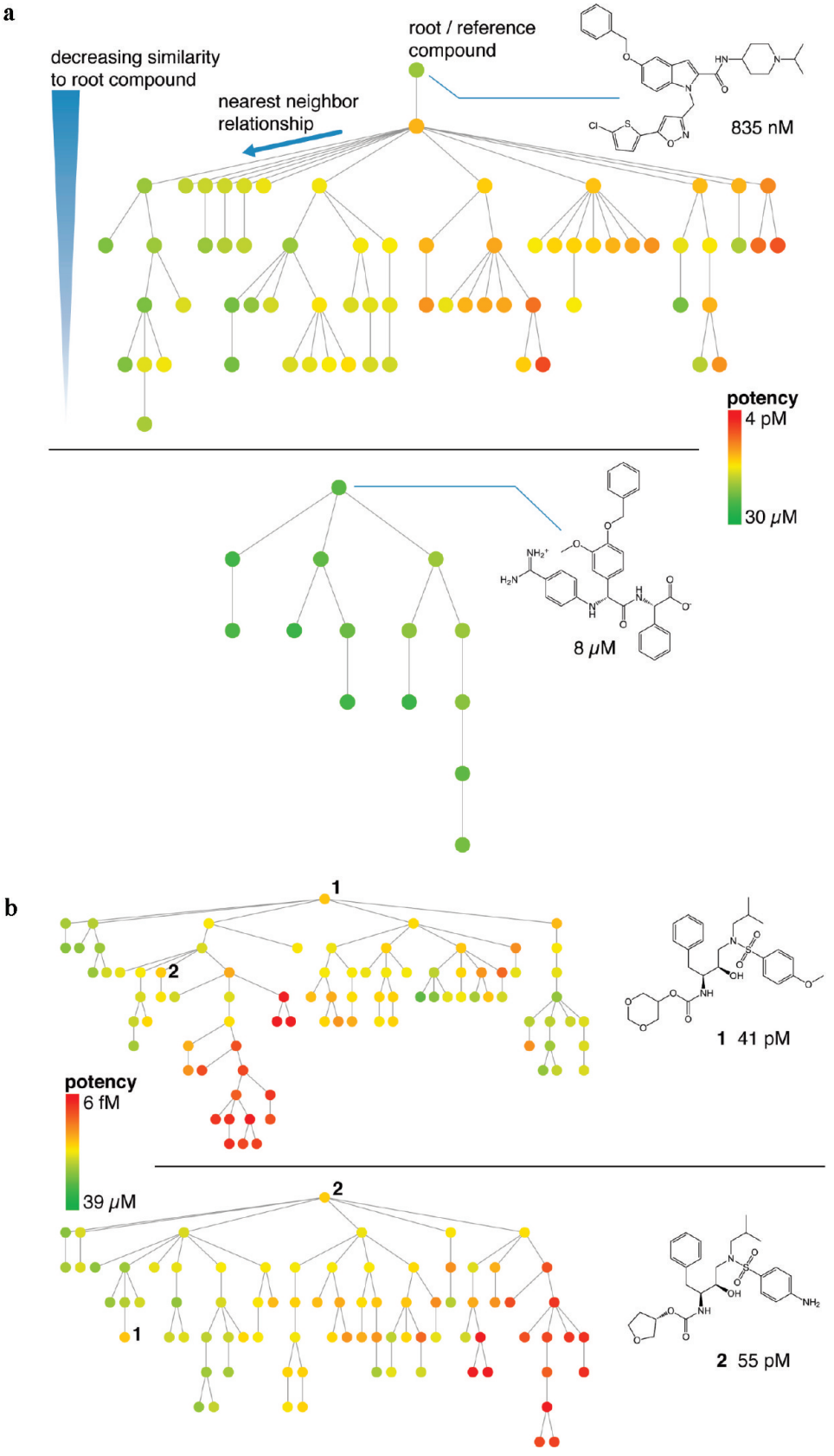

Figure 7. Similarity−potency trees. (a) Two annotated SPTs are shown for a set of factor Xa inhibitors. Nodes represent compounds, node color reflects compound potency, and edges connect structural nearest neighbors. From the top to the bottom, similarity to the root (reference) compound decreases. (b) Two SPTs for a set of HIV protease inhibitors illustrate the effects of reference compound selection. Both trees contain nearly identical compound subsets. The respective root nodes for both trees are indicated, and their structures are shown on the right. The figure has been adapted from ref 32.

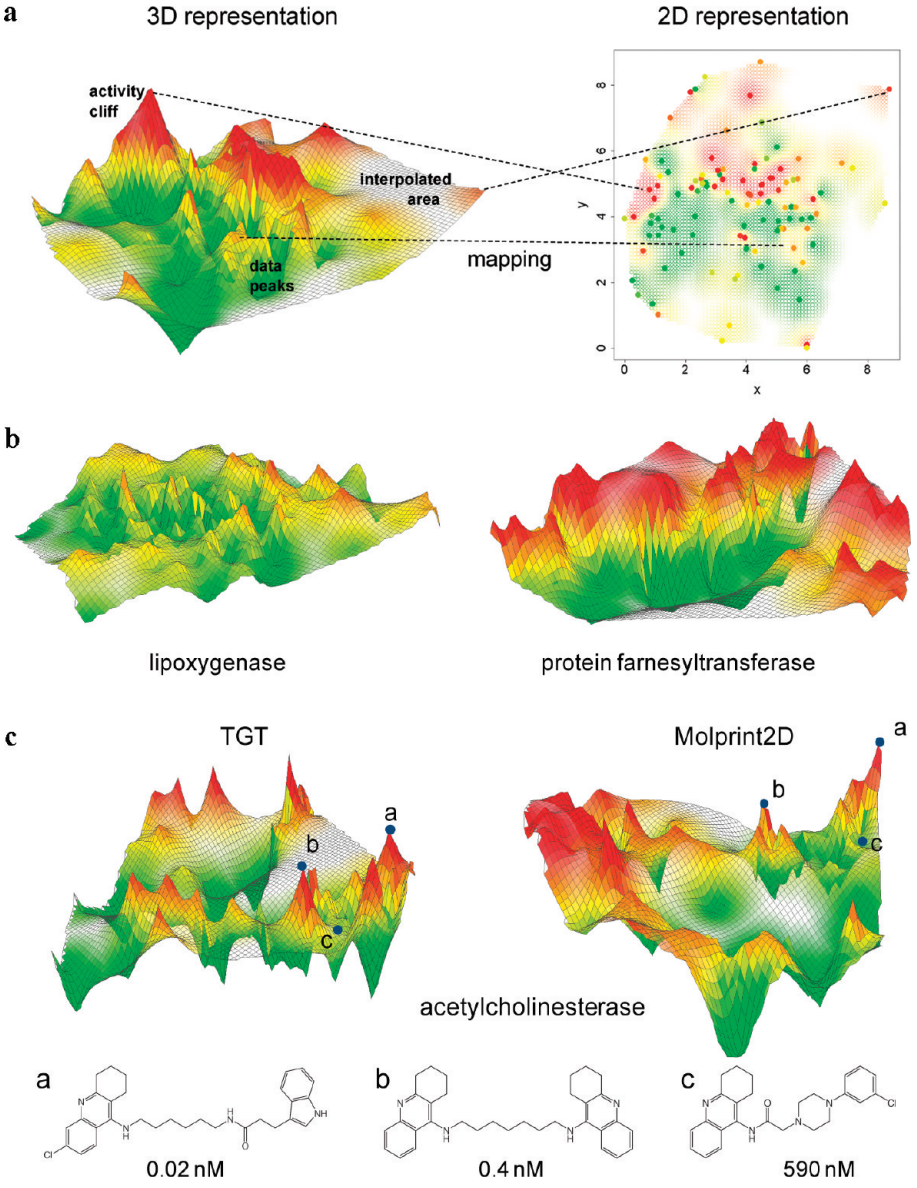

Figure 8. Detailed 3D activity landscapes. (a) For a set of 112 acetylcholinesterase inhibitors represented as MACCS fingerprints, a detailed 3D activity landscape is shown and compared to the underlying 2D representation obtained by dimension reduction. Distinct regions of the 3D landscape are annotated and further discussed in the text. (b) 3D activity landscapes based on Molprint2D fingerprint representations are shown for sets of 252 and 146 inhibitors of lipoxygenase and protein farnesyltransferase, respectively. The activity landscape for lipoxygenase is characterized by overall SAR continuity, whereas the landscape for protein farnesyltransferase reveals a high degree of SAR discontinuity contained in this data set. (c) For the set of acetylcholinesterase inhibitors used in (a), landscapes are generated with alternative TGT and Molprint2D fingerprint representations. The positions of three compounds are mapped and their potency is reported. Depending on the molecular representation, the topology of the landscapes changes and the compounds form, or do not form, activity cliffs. The figure has been adapted from ref 33.

compounds active against two targets, activity cliff markers and compounds that do not form activity cliffs may or may not form selectivity cliffs. Accordingly, in the context of multi-target SARs, variable relationships between activity and selectivity cliffs exist.

Compound-Centric Landscape Views

In NSGs, compound similarity relationships are accounted for by building a global network structure from individual pairwise compound comparisons. Different from this approach, compound-centric landscape views can also be generated that define a single compound as a structural reference so that it is possible to describe the "coordinates" of all other compounds in a data set relative to this reference point in chemical space. An example for an SAR analysis method that relies on a compound-centric data view is provided by similarity-potency trees (SPTs).[32] Like NSGs, SPTs are graph-based data structures, i.e. colored nodes represent compounds and their potency values and edges indicate similarity relationships (node scaling is not applied here). However, SPTs focus on local SARs and a unique feature is that they represent a compound hierarchy as a tree based on similarity to the reference compound (Figure 7a). The reference compound at the top of the hierarchy forms the root of the tree. From top to bottom, the similarity of the compounds to the reference compound decreases until a minimum similarity threshold is reached. Edges connect compounds that are structurally most similar to each other (i.e., nearest neighbors). Accordingly,

gradual structural modifications of the reference compound and their effect on potency are visualized. To simplify the identification of SAR trends, children of the same parent compound are sorted by increasing potency from left to right.

A question associated with a compound-centric view of an activity landscape is how to best select a compound as the reference. In SPT analysis, one attempts to identify trees that simplify SAR analysis. To avoid the a priori definition of an SAR model, each compound in a given data set is selected once as a reference to calculate an SPT. After this complete set of SPTs has been generated, the data structures are ranked based on their predefined SAR information content.[32] By systematically exploring all possible reference compounds, no SAR information is lost. In Figure 7a, two exemplary SPTs of a set of factor Xa inhibitors are shown that display different levels of SAR information. The tree at the top contains many compounds exceeding the minimum similarity threshold to the reference molecule, and their potency values are distributed following a fairly regular pattern. Most compounds have multiple children with different potency. These children can be readily compared to elucidate structural modifications that influence the compound potency. By contrast, the tree at the bottom in Figure 7a conveys only little SAR information. It consists of a small number of compounds that all have low and nearly invariant potency values.

The effects of reference compound selection on the interpretability of SPTs are illustrated in Figure 7b. Both of the trees that are shown consist of nearly identical compound sets. But the selection of compounds **1** and **2** as respective root nodes results in different potency distributions. The tree shown at the top (rooted at compound **1**) does not exhibit obvious patterns in the distribution of potency values. By contrast, in the tree at the bottom (rooted at compound **2**), potency tends to increase from left to right on each tree level, although the compounds are not connected to the same parent. This indicates that compounds with high potency tend to have highly potent children. Because potency levels are generally retained when moving down along this SPT, SAR analysis becomes feasible. Thus, SPTs are designed to focus SAR exploration on local regions of an activity landscape with evident SAR patterns and to organize compounds in these regions in an interpretable manner.

Detailed 3D Activity Landscapes

The idealized schematic 3D landscape views discussed above provide an intuitive basis to rationalize principal SAR phenotypes. Of course, one would also be interested in generating such 3D activity landscapes for actual data sets and study their topology. In a recent study, this has been attempted.[33] Following this approach, coordinate-free chemical reference spaces are generated for compound data sets based on pairwise distances between fingerprint representations. Then 2D projections of these reference spaces are calculated using multidimensional scaling[34] as a dimension reduction technique and compound positions are mapped onto the $x-y$ plane of a coordinate system. Compound potency values are reported on the z-axis. To obtain a contiguous activity surface from sparsely distributed compound potency data, interpolation functions are applied. Figure 8a shows an exemplary 3D landscape for a set of acetylcholinesterase inhibitors and the 2D projection from which it is derived. As the molecular representation, MACCS structural keys[35] are used. The activity surface is color-coded according to surface elevation by applying a continuous (green to red) spectrum. Interpolated surface area is displayed in white. In this case, closely corresponding 2D and 3D representations of an activity landscape can be compared and the positions of activity cliffs and other compounds can be mapped. In comparison to idealized 3D landscape models, 3D activity landscapes of compound data sets typically contain many more peaks that are often, however, not significant cliffs. Rather, small peaks are a consequence of dense local data sampling and are hence termed "data peaks" (Figure 8a). Because of the surface elevation-dependent color scheme, activity cliffs of significant magnitude appear in red. As shown in Figure 8a, activity cliffs can also be identified in the 2D projection of chemical reference space at intersections of red and green areas, but the 3D landscape adds topological information about cliffs. For example, it differentiates steep peaks from activity plateaus that are formed by multiple compounds. From 3D landscape views of different data sets, SAR characteristics can indeed be deduced. For example, Figure 8b shows the comparison of activity landscapes of sets of lipoxygenase and farnesyltransferase inhibitors that are characterized by strong SAR continuity and discontinuity, respectively. In this case, an atom environment fingerprint (Molprint2D)[36] is used to represent the test compounds, and the activity landscapes are produced using a common reference coordinate system (making them directly comparable). By comparing these landscapes, differences in global SAR character become immediately apparent. The lipoxygenase inhibitor landscape has a somewhat rugged surface due to the presence of many data peaks (see above) but is overall gently sloped and does not display activity cliffs. By contrast, the farnesyltransferase inhibitor landscape is characterized by the presence of many activity cliffs.

Considering the general similarity caveat, we can also study the effects of using different molecular representations on the topology of 3D activity landscapes. Figure 8c shows a comparison of two landscapes calculated for the acetylcholinesterase inhibitor set using alternative molecular representations, i.e. Molprint2D and a 2D pharmacophore fingerprint (TGT).[37] These two landscapes differ significantly, which illustrates the strong influence of chosen molecular representations (and chemical reference spaces) on the nature of calculated similarity relationships and ensuing SARs. For example, as illustrated in Figure 8c, activity cliffs that are apparent in one landscape might be substantially altered, or even absent, in another, and landscape topology might vary greatly. This is not an intrinsic limitation of 3D landscape modeling, but rather a phenomenon that affects all types of computational SAR analysis. In fact, the influence of chosen molecular representations on landscape topology can be utilized as a diagnostic tool to visualize and assess how different representations alter SAR features of compound data sets. For this purpose, alternative landscapes can be readily computed and compared.[33] For example, if a compound data set should be subjected to QSAR modeling, it would make sense to evaluate alternative activity landscapes generated using different types of descriptors and select representations that induce a higher degree of SAR continuity than others. Although the 3D landscapes discussed herein are based on fingerprint representations, they can also be generated from real-valued molecular descriptor spaces using dimension reduction techniques such as principal component analysis.[26]

SAR Determinants in Analogue Series

The landscape views described so far are applicable to large data sets of different composition. However, in advanced stages of compound design, optimization efforts are typically focused on a small part of activity landscapes populated by analogues of a single chemotype. Therefore, for the study of analogue series, different types of activity landscape models can be envisioned. For example, a hierarchical tree-like data structure termed a combinatorial analogue graph (CAG) has been introduced to systematically organize analogue series according to substitution patterns (on the basis of R-group decomposition) and assess the degree of SAR discontinuity that substitutions at defined sites, and their combination, introduce (Figure 9a).[38] For this purpose, it is also meaningful to depart from the assessment of whole-molecule similarity utilized in other landscape representations and evaluate local similarity focusing on R-group patterns. To these ends, pharmacophore edit distances have recently been applied to compare analogues exclusively by the similarity of their substituents.[39] In CAGs, nodes correspond to compound subsets with substitutions at defined sites or site combinations. Node labels identify the substitution sites at which the analogue pairs of a given node (subset) differ. Edges connect subsets that share modifications at one or more substitution sites. Nodes are further annotated by their normalized discontinuity scores that are also reflected by the color code using a continuous color spectrum from green (low discontinuity) to red (high discontinuity). At the top of Figure 9a, a CAG is shown for six analogues active against cyclin-dependent kinase 4 having three variable substitution sites that illustrates the assignment of compound pairs to corresponding subsets. As can be seen, the discontinuity score is highest for pairs of compounds that differ at site 1. Hence, this site constitutes a so-called "SAR hotspot". Here, changes are most likely to introduce SAR discontinuity and produce compounds with large differences in potency. Furthermore, the modification of site 1 in combination with substituent exchanges at other positions also generates SAR discontinuity, as reflected by the red nodes 1−2 and 1−3. As analogue sets grow in size, more substituent positions can be explored, as shown for a series of 36 cytochrome P450 3A4 inhibitors at the bottom of Figure 9a. This CAG representation identifies the substitution sites 1, 2, and 6 as SAR hotspots. Furthermore, it reveals "SAR holes", i.e. combinations of substitution sites, that have not yet been explored. Hence, useful suggestions which compounds to synthesize next and how to complement a current series can be derived from considering relationships between SAR holes and hotspots.

CAGs have also been applied to study multitarget SARs.[39] Figure 9b shows corresponding CAGs for a series of 18 inhibitors with three variable substitution sites that are active against the four serine proteases factor Xa, thrombin, urokinase, and trypsin. It is evident that factor Xa and thrombin show very similar patterns of SAR discontinuity. In both cases, site 1 is mainly responsible for the formation of activity cliffs whereas for urokinase, sites 2 and 3 are SAR hotspots. By contrast, the CAG representation for trypsin is characterized by overall low SAR discontinuity. SAR hotspots that are unique to one or few targets provide the opportunity to selectively alter the potency of analogues for individual targets. Taken together, CAGs integrate information about potency changes with a local view on compound similarity and provide immediate access to substitution sites most relevant for changes in compound potency or selectivity. They provide a compound organization and view of small sections of activity landscapes centered on a given chemotype.

Activity Landscape Analysis for Medicinal Chemistry

Herein we have discussed alternative ways to conceptualize and represent activity landscapes for compound data sets. Implementations of most of the activity landscape methods discussed herein are freely available to the scientific community (as specified in the original publications). All activity landscape software tools developed in our laboratory can be obtained via the download section of our Web site (http://www.lifescienceinformatics.uni-bonn.de), some also in integrated form as part of the SARANEA program.[40] The interpretability of landscape models is generally supported by interactively associating compound nodes with 2D depictions of compound structures. Common features of all landscape models include that they are based on systematic comparisons of compound similarity and potency relationships and that they are designed to reveal SAR characteristics. However, details and specific purposes of the activity landscape models presented herein differ considerably and there might well be individual preferences for one or the other model. Activity landscapes are only one of many ways to model and analyze SARs. From our point of view, particularly important aspects of the activity landscape approach include that: (i) intuitive graphical access to SAR characteristics is provided at varying levels of complexity, (ii) global and local SAR features can be related to each other for (iii) compound data sets of increasing size, (iv) regardless of their structural homogeneity, (v) it provides a basis for chemical interpretation, but does not attempt to replace it.

How might such activity landscape representations aid in practical medicinal chemistry applications? A characteristic feature of landscape modeling is that it is not predictive but descriptive in nature. Activity landscapes of data sets of different composition help to view SAR information in context. One can quickly understand to what extent a data set contains SAR information. If it does, it is possible to focus on compound subsets displaying characteristic local SAR features and spot prominent activity cliffs. As such, landscape views aid in compound selection. Supported by numerical analysis functions, compound subsets corresponding to regions of high SAR discontinuity (SARI) or activity cliff patterns (SALI) can easily be identified. Although there is usually strong emphasis on the identification of activity cliffs in landscape analysis, cliffs do not necessarily provide interpretable SAR information, as we have discussed. Hence, it is equally important to inspect different compound subsets in regions of apparent SAR continuity and discontinuity and search for interpretable SAR information. Landscape representations such as NSGs can be mined for compound pathways that connect continuous and discontinuous SAR regions in data sets, which further aids in the search for interpretable information. It is also possible to identify compound subsets in activity landscapes that have already been thoroughly explored without revealing significant SAR discontinuity, which should help to deprioritize "flat" SARs. Taken together, graphical access to SAR information at the level of whole data sets and compound subsets, the evaluation of SAR patterns, and the identification of key compounds that induce local SARs are major aspects of activity landscape analysis for medicinal chemistry applications.

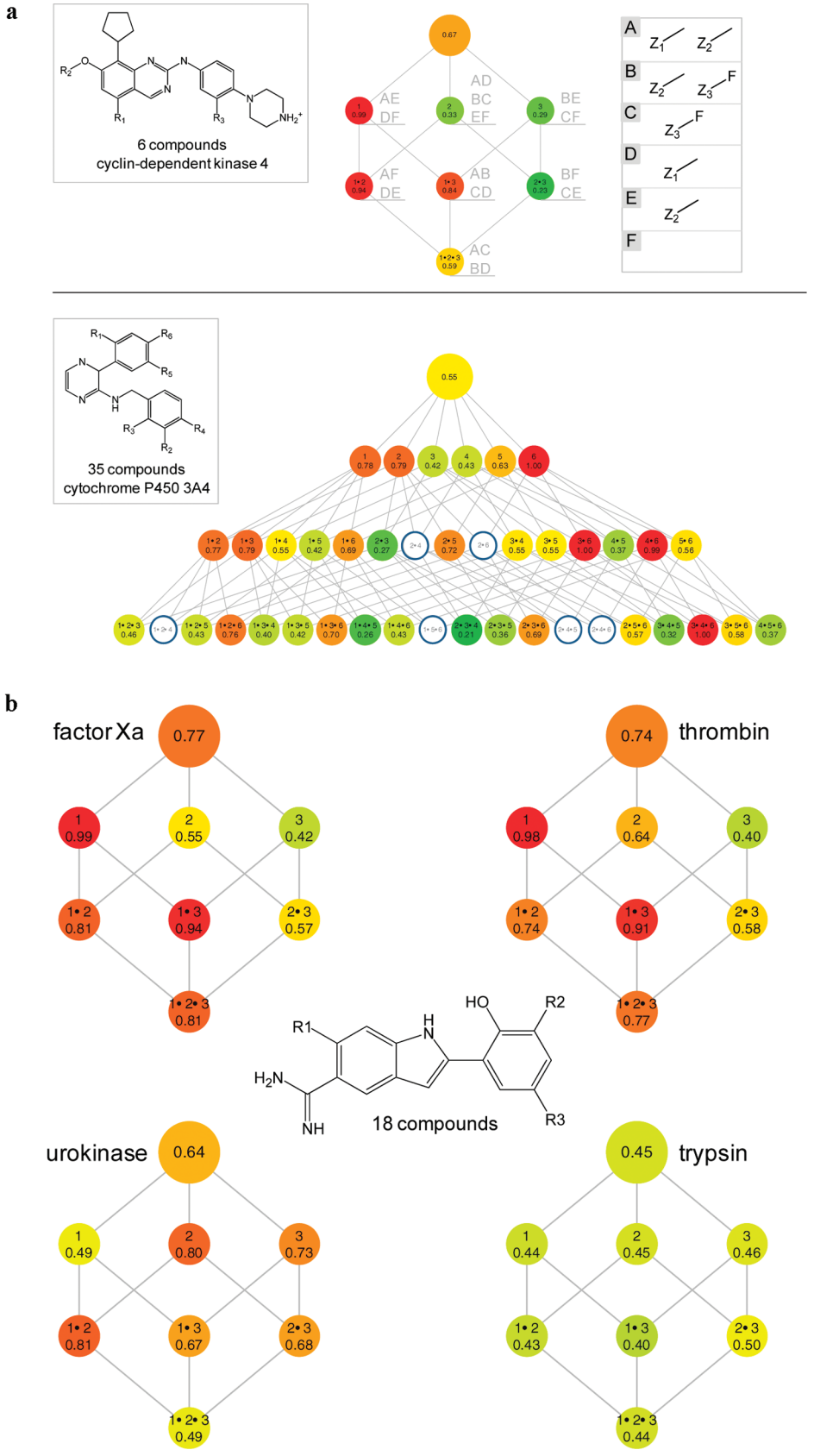

Figure 9. Combinatorial analogue graphs. (a) CAG representations for six and 35 analogues active against the protein targets cyclin-dependent kinase 4 (CDK4) and cytochrome P450 3A4 (CYP P450 3A4), respectively, are shown. For the CDK4 inhibitors, substituents of the individual compounds and the assignment of compound pairs to corresponding nodes are reported. SAR holes in the CAG representation for CYP P450 3A4 are circled in blue. (b) For a series of 18 compounds active against the serine proteases factor Xa, thrombin, urokinase, and trypsin, the four corresponding CAG representations are shown. Different SAR hotspots are detected that can be exploited in the design of potent and selective analogues.

Conclusions and Outlook

Activity landscape representations are attractive tools to access SAR information contained in compound data sets of any source. Landscape analysis is best applied to moderately sized compound data sets. If data sets are very large, network-based landscape views become difficult to analyze graphically. If data sets are very small (i.e., containing only tens of compounds), landscape views are not required to analyze SAR information. However, even for data sets containing only on the order of 50−100 compounds, landscape models can be readily generated and are often very informative. The description of activity landscapes principally relies on systematic comparisons of compound similarity and potency and graphical representations of the results. Although the concept of activity landscapes has been introduced already a number of years ago, until recently only very few studies describing landscape models had been reported. However, catalyzed by the introduction of numerical SAR analysis functions that depart from the classical QSAR paradigm and the introduction of molecular network representations, several 2D and 3D activity landscape models have recently been reported. These models present activity landscapes in rather different ways and provide alternative view points to intuitively access and compare global and local SAR features. However, in particular the modeling of 3D landscapes is still in its infancy and we expect to see increasing efforts in the near future to derive detailed landscapes for alternative molecular representations that utilize different methodological frameworks. It is conceivable that such landscape models will be used as diagnostics to test the suitability of alternative chemical reference spaces to capture SAR information. We also anticipate that interactive activity landscape modeling will become increasingly popular for retrospective SAR exploration of historically grown and increasingly large sets of active compounds. Landscape modeling should help to monitor the evolution of compound data sets in pharmaceutical settings.

Acknowledgment. We thank Lisa Peltason for discussions and help with illustrations.

Biographies

Anne Mai Wassermann studied Molecular Biomedicine at the University of Bonn, Germany. In 2008, she joined the Department of Life Science Informatics at the University of Bonn headed by Prof. Jürgen Bajorath for her masters project, where she worked on machine-learning methods for virtual screening. Anne Mai is now in the second year of her Ph.D. studies, and her current research interests are in the analysis of multitarget structure−activity relationships.

Mathias Wawer studied Molecular Biomedicine at the University of Bonn, Germany, where he joined the Department of Life Science Informatics headed by Prof. Jürgen Bajorath for his masters project in 2008. Currently, he works on the development of graphical methods for the systematic computational analysis of structure−activity relationships. His Ph.D. work is supported by the Lead Discovery Department of Boehringer Ingelheim Pharma, Biberach, Germany.

Jürgen Bajorath is Professor and Chair of Life Science Informatics at the University of Bonn. He is also an Affiliate Professor in the Department of Biological Structure at the University of Washington, Seattle. His research interests include early phase drug discovery and the development of computational methods for molecular similarity analysis and the systematic exploration of structure−activity relationships. For more details, please see: https://www.lifescienceinformatics.uni-bonn.de.

References

(1) Lajiness, M. S.; Maggiora, G. M.; Shanmugasundaram, V. Assessment of the Consistency of Medicinal Chemists in Reviewing Sets of Compounds. *J. Med. Chem.* **2004**, *47*, 4891–4896.
(2) Nicholls, A. What do we know and when do we know it? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 239–255.
(3) Ahlberg, C. Visual Exploration of HTS Databases: Bridging the Gap between Chemistry and Biology. *Drug Discovery Today* **1999**, *4*, 370–376.
(4) Roberts, G.; Myatt, G. J.; Johnson, W. P.; Cross, K. P.; Blower, P. E., Jr. Lead Scope: Software for Exploring Large Sets of Screening Data. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1302–1314.
(5) Agrafiotis, D.; Shemanarev, M.; Connolly, P.; Farnum, M.; Lobanov, V. SAR Maps: A New SAR Visualization Technique for Medicinal Chemists. *J. Med. Chem.* **2007**, *50*, 5926–5937.
(6) Manallack, D. T.; Ellis, D. D.; Livingstone, D. J. Analysis of Linear and Nonlinear QSAR Data using Neural Networks. *J. Med. Chem.* **1994**, *37*, 3758–3767.
(7) Esposito, E. X.; Hopfinger, A. J.; Madura, J. D. Methods for Applying the Quantitative Structure−Activity Relationship Paradigm. *Methods Mol. Biol.* **2004**, *275*, 131–214.
(8) Leach, A. R.; Gillet, V., J.; Lewis, R. A.; Taylor, R. Three-dimensional Pharmacophore Methods in Drug Discovery. *J. Med. Chem.* **2010**, *53*, 539–558.
(9) Geppert, H.; Vogt, M.; Bajorath, J. Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205–216.
(10) 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.
(11) Dobson, C. M. Chemical Space and Biology. *Nature* **2004**, *432*, 824–828.
(12) Lipinski, C.; Hopkins, A. Navigating Chemical Space for Biology and Medicine. *Nature* **2004**, *432*, 855–861.
(13) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior—A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996**, *39*, 3049–3059.
(14) 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.
(15) Peltason, L.; Bajorath, J. SAR Index: Quantifying the Nature of Structure−Activity Relationships. *J. Med. Chem.* **2007**, *50*, 5571–5578.
(16) Brown, J.; Jacoby, E. On scaffolds and hopping in medicinal chemistry. *Mini. Rev. Med. Chem.* **2006**, *6*, 1217–1229.
(17) Lajiness, M. Evaluation of the Performance of Dissimilarity Selection Methodology. In *QSAR: Rational Approaches to the Design of Bioactive Compounds*; Silipo, C., Vittoria, A., Eds.; Elsevier: Amsterdam, 1991; pp 201−204.
(18) Maggiora, G. M.; Shanmugasundaram, V.; Lajiness, M. S.; Doman, T. N.; Schulz, M. W. A Practical Strategy for Directed Compound Acquisition. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2005; pp 317−332.
(19) Peltason, L.; Bajorath, J. Systematic Computational Analysis of Structure−Activity Relationships: Concepts, Challenges and Recent Advances. *Future Med. Chem.* **2009**, *1*, 451–466.
(20) Maggiora, G. M. On Outliers and Activity Cliffs—Why QSAR often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535–1535.
(21) Johnson, M. A., Maggiora, G. M., Eds. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, 1990.
(22) 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.
(23) Wawer, M.; Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Structure−Activity Relationship Anatomy by Network-like Similarity Graphs and Local Structure−Activity Relationship Indices. *J. Med. Chem.* **2008**, *51*, 6075–6084.
(24) Guha, R.; Van Drie, J. H. Structure−Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model* **2008**, *48*, 646–658.
(25) 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.
(26) Guha, R. The Ups and Downs of Structure−Activity Landscapes. *Methods in Molecular Biology*; Bajorath, J, Ed.; Springer: New York, 672.

(27) Wawer, M.; Peltason, L.; Bajorath, J. Elucidation of Structure−Activity Relationship Pathways in Biological Screening Data. *J. Med. Chem.* **2009**, *52*, 1075–1080.
(28) Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2005; pp 271−285.
(29) Wassermann, A. M.; Bajorath, J. Chemical Substitutions That Introduce Activity Cliffs across Different Compound Classes and Biological Targets. *J. Chem. Inf. Model.* **2010**, *50*, 1248−1256.
(30) Hu, Y.; Bajorath, J. Molecular Scaffolds with High Propensity to Form Multitarget Activity Cliffs. *J. Chem. Inf. Model.* **2010**, *50*, 500–510.
(31) 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.
(32) 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 in press.
(33) Peltason, L.; Iyer, P.; Bajorath, J. Rationalizing Three-Dimensional Activity Landscapes and the Influence of Molecular Representations on Landscape Topology and the Formation of Activity Cliffs. *J. Chem. Inf. Model* **2010**, *50*, 1021–1033.
(34) Borg, I.; Groenen, P .J. F. *Modern Multidimensional Scaling. Theory and Applications*, 2nd ed.; Springer: New York, 2005.
(35) *MACCS Structural Keys*; Symyx Software: San Ramon, CA, 2002.
(36) *Molecular Operating Environment* (*MOE*); Chemical Computing Group, Inc.: Montreal, Quebec, Canada, 2007.
(37) 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.
(38) Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Exploration of Structure−Activity Relationship Determinants in Analogue Series. *J. Med. Chem.* **2009**, *52*, 3212–3224.
(39) Wassermann, A. M.; Peltason, L.; Bajorath, J. Computational Analysis of Multitarget Structure−Activity Relationships to Derive Preference Orders for Chemical Modifications toward Target Selectivity. *ChemMedChem* **2010**, *5*, 847–858.
(40) Lounkine, E.; Wawer, M.; Wassermann, A. M.; Bajorath, J. SARANEA: A Freely Available Program To Mine Structure−Activity and Structure−Selectivity Relationship Information in Compound Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 68–78.