Structure−Activity Relationship Anatomy by Network-like Similarity Graphs and Local Structure−Activity Relationship Indices

Mathias Wawer,[†,§] Lisa Peltason,[†,§] Nils Weskamp,[‡] Andreas Teckentrup,[‡] 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, and Department of Lead Discovery, Boehringer Ingelheim Pharma GmbH & Co. KG, D-88397 Biberach/Riss, Germany

Received July 14, 2008

The study of structure−activity relationships (SARs) of small molecules is of fundamental importance in medicinal chemistry and drug design. Here, we introduce an approach that combines the analysis of similarity-based molecular networks and SAR index distributions to identify multiple SAR components present within sets of active compounds. Different compound classes produce molecular networks of distinct topology. Subsets of compounds related by different local SARs are often organized in small communities in networks annotated with potency information. Many local SAR communities are not isolated but connected by chemical bridges, i.e., similar molecules occurring in different local SAR contexts. The analysis makes it possible to relate local and global SAR features to each other and identify key compounds that are major determinants of SAR characteristics. In many instances, such compounds represent start and end points of chemical optimization pathways and aid in the selection of other candidates from their communities.

Introduction

The analysis of structure−activity relationships of small molecules is a major focal point in lead optimization and drug design.[1] Typically, SARs[a] are explored on a case-by-case basis, given a specific target and series of active compounds. By contrast, systematic analyses of SARs over many compound classes have thus far been rarely reported. It is widely appreciated that relationships between molecular structure and biological properties are often highly complex.[1−3] In many instances, small chemical modifications of active compounds can have dramatic positive or negative effects on potency and/or selectivity, a situation that is very familiar to medicinal chemists. In other cases, chemical modifications of hits or leads cause only small or gradual changes in activity. This situation might even give rise to "flat SARs" where significant chemical modifications of active compounds change biological activity only very little. This SAR behavior presents a particularly complicated scenario for medicinal chemists because it might remain unclear whether compound series displaying such characteristics can be further optimized.

The nature of SARs is ultimately determined by biological response characteristics of chemical changes (i.e., by the way chemical modifications affect specific ligand−target interactions). These characteristics can be conceptualized as topological maps of potency distributions in chemical space, so-called "activity landscapes".[2] In these topology maps, compound potency is added as a third dimension to a 2D projection of chemical space.[3] If similar compounds have significantly different potency, a small move within the *xy*-plane is accompanied by a large change in *z*-direction, and the resulting topology is rugged or canyon-like. By contrast, chemically different compounds having similar potency produce a rolling hill-like topology. Accordingly, SARs characterized by rugged landscapes have been termed "discontinuous" SARs and those characterized by rolling hill-like topology "continuous" SARs.[3] Canyon-like local topologies correspond to "activity cliffs"[2] where small changes in structure lead to significant changes in compound potency. Activity cliffs provide the basis for lead optimization, whereas rolling hill topologies are a prerequisite for the successful application of whole-molecule similarity methods that aim to identify structurally diverse compounds having similar activity.[3] These similarity methods have their foundation in the similarity−property principle[4] stating that overall similar molecules (however "similarity" is defined) should also have similar activity. Thus, it assumes the presence of continuous SARs.

The concept of activity landscapes makes it possible to characterize different types of SARs. For this purpose, structure−activity similarity (SAS) maps were originally designed to compare compounds in a pairwise manner and relate potency differences and similarity to each other.[5] SAS maps organize sets of active compounds into groups with different potency−similarity relationships. Other SAR maps were also developed in order to represent large compound data sets and identify SARs they might contain.[6] Furthermore, in order to distinguish between principal SAR features, a systematic analysis was carried out on different series of ligands in complex crystal structures, relating 2D similarity, 3D (binding mode) similarity, and compound potency to each other.[7] This analysis demonstrated that many SAR landscapes are heterogeneous in nature because they contain both continuous and discontinuous regions. In other words, the underlying activity landscapes consist of activity cliffs that are separated by rolling hills. Targeting such heterogeneous SARs is thought to be particularly attractive for medicinal chemistry efforts because it is likely that structurally diverse

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

[†] Rheinische Friedrich-Wilhelms-Universität.

[§] These authors contributed equally to this work.

[‡] Boehringer Ingelheim Pharma GmbH & Co. KG.

[a] Abbreviations: NSG, network-like similarity graph; SALI, structure−activity landscape index; SAR, structure−activity relationship; SARI, structure−activity relationship index; SAS, structure−activity similarity; Tc, Tanimoto coefficient; COX, cyclooxygenase-2; FAR, protein farnesyltransferase; FXA, coagulation factor Xa; LIP, lipoxygenase; SQA, squalene synthase; THR, thrombin.

10.1021/jm800867g CCC: $40.75
Published on Web 09/18/2008

Table 1. Enzyme Inhibitor Classes for SAR Analysis[a]

		no. of	MACCS Tc			potency			global scores			global
class	activity	compds	min	max	av	min	max	av	cont	disc	SARI	SAR type
LIP	lipoxygenase inhibitors	252	0.02	1.00	0.36	4.00	9.00	6.61	0.99	0.04	0.97	continuous
COX	cyclooxygenase-2 inhibitors	149	0.05	1.00	0.45	4.30	10.05	6.92	0.74	0.21	0.77	continuous
FXA	factor Xa inhibitors	152	0.11	1.00	0.50	4.52	11.15	7.97	0.30	0.27	0.52	heterogeneous-constrained
FAR	protein farnesyl transferase inhibitors	146	0.01	1.00	0.45	3.52	10.44	7.82	0.58	0.71	0.44	heterogeneous-relaxed
SQA	squalene synthase inhibitors	71	0.08	1.00	0.44	3.30	10.15	7.50	0.79	0.99	0.40	heterogeneous-relaxed
THR	thrombin inhibitors	172	0.14	1.00	0.55	4.52	11.72	8.15	0.08	0.67	0.21	discontinuous

[a] Compound activity classes assembled from the MDDR are reported together with their potency (pK_i or pIC$_{50}$ values) and similarity distributions and global SARI scores. "av" stands for average and "cont" and "disc" stand for continuity and discontinuity, respectively. Global SAR types are reported according to the SARI classification scheme.[8]

active compounds can be identified (on rolling hill areas) and also optimized (if they map to the vicinity of activity cliffs).[7]

Such qualitative insights into global SAR features were complemented and further extended by the introduction of numerical functions to characterize SARs in a quantitative manner.[8,9] The SAR index (SARI) combines continuity and discontinuity scores that are independently calculated from 2D compound similarity and potency.[8] For various compound activity classes, global SARI scores were calculated that indicate the presence of activity cliffs, gently sloped activity landscapes, or their combination. SARI scoring of many different activity classes has shown that SARs are often heterogeneous in nature but has also identified purely continuous and discontinuous SARs.[8] Heterogeneous SARs were further classified either as "heterogeneous-relaxed" where different compound series representing continuous or discontinuous SARs are found within an activity class or, alternatively, as "heterogeneous-constrained" where structural variations with little consequences for compound potency are observed within the regional constraints of an activity cliff. With the structure–activity landscape index (SALI), another function to quantify SARs was introduced. Different from SARI, SALI was designed to only detect activity cliffs by pairwise compound comparison. SALI score relationships are graphically illustrated by representing compounds as nodes in a graph and connecting them with edges if the pairwise SALI score is above a predefined threshold value.[9] This type of illustration makes it possible to follow compound modifications that are associated with activity cliffs.

SARI scoring permits a global assessment of SAR features, and both SARI and SALI analysis enable the identification of activity cliffs. However, these functions cannot be applied to study multiple SAR features contained within a set of active compounds at the level of individual molecules and their relationships to others. In order to dissect SAR landscapes and analyze multiple SAR components of compound classes with different SAR behavior, we introduce network-like similarity graphs (NSG) and a "local" SARI score variant that is capable of accounting for SAR contributions from individual compounds. In computational medicinal chemistry, network representations have previously been used to represent ligand–target relationships[10,11] or relationships between different classes of drug molecules.[12] In the latter study, it was found that drug networks generally have small-world character, which means that drug molecules are predominantly organized in communities of closely related compounds. Furthermore, in another recent study, known drug–target interactions were systematically mapped onto protein–protein interaction networks in order to facilitate quantitative network analysis and characterize drugs directed at novel targets.[13]

Here, we introduce and utilize NSGs and local SARI scoring to describe different SAR features that coexist in sets of active compounds. This type of "SAR anatomy" makes it possible to better understand how local SAR characteristics in compound activity classes are related to each other and identify individual active molecules that are SAR determinants.

Materials and Methods

SARI. We calculate the global SARI score as described previously[8] for sets of active compounds and subsets obtained by similarity-based clustering. The SARI scoring scheme has been introduced to determine continuous and discontinuous SAR components in a given set of active compounds.[8] It is composed of two separately calculated scores, the continuity and discontinuity score. The initially computed "raw" scores are normalized with respect to a reference panel of 13 compound activity classes from the molecular drug data report (MDDR).[14] The continuity score quantifies the composition of smooth regions within an activity landscape, reflected by active compounds with increasing structural diversity but similar biological activity. It is derived from the weighted mean of pairwise compound similarity calculated using MACCS keys[15] and the Tanimoto coefficient (Tc).[16] The weights combine the potency values of each compound and the difference in potency between them. Here, we use a modified version of the original implementation that further emphasizes global diversity of a compound set. The modified raw score is defined as

$$\mathrm{raw}_{\mathrm{cont}} = \frac{\sum_{\{i,j|i>j\}} w_{ij} \frac{1}{1 + \mathrm{sim}(i,j)}}{\sum_{\{i,j|i>j\}} w_{ij}}, \qquad w_{ij} = \frac{\mathrm{pot}(i)\mathrm{pot}(j)}{1 + \mathrm{potdiff}(i,j)}$$

where pot(i) gives the potency value of compound i (as pK_i or pIC$_{50}$ value), potdiff(i,j) denotes the absolute potency difference, and sim(i,j) denotes the MACCS Tc similarity between compounds i and j. Hence, the continuity score measures the global diversity in a set of compounds that are active against the same target. The weighting scheme ensures that compound pairs where both compounds have similarly high potency contribute more to the score than compound pairs with low potency and/or large potency differences, taking into account that SAR continuity is primarily characterized by small potency differences among structurally diverse compounds.

The raw discontinuity score is designed to detect and quantify the presence of activity cliffs in a compound set. Consequently, it is defined as the average potency difference between ligand pairs exceeding a predefined similarity threshold multiplied by the pairwise ligand similarity. We set this similarity threshold to 0.65 (i.e., selecting relatively similar compounds) in order to focus the analysis on significant activity cliffs. Multiplication with similarity puts high emphasis on potency differences among

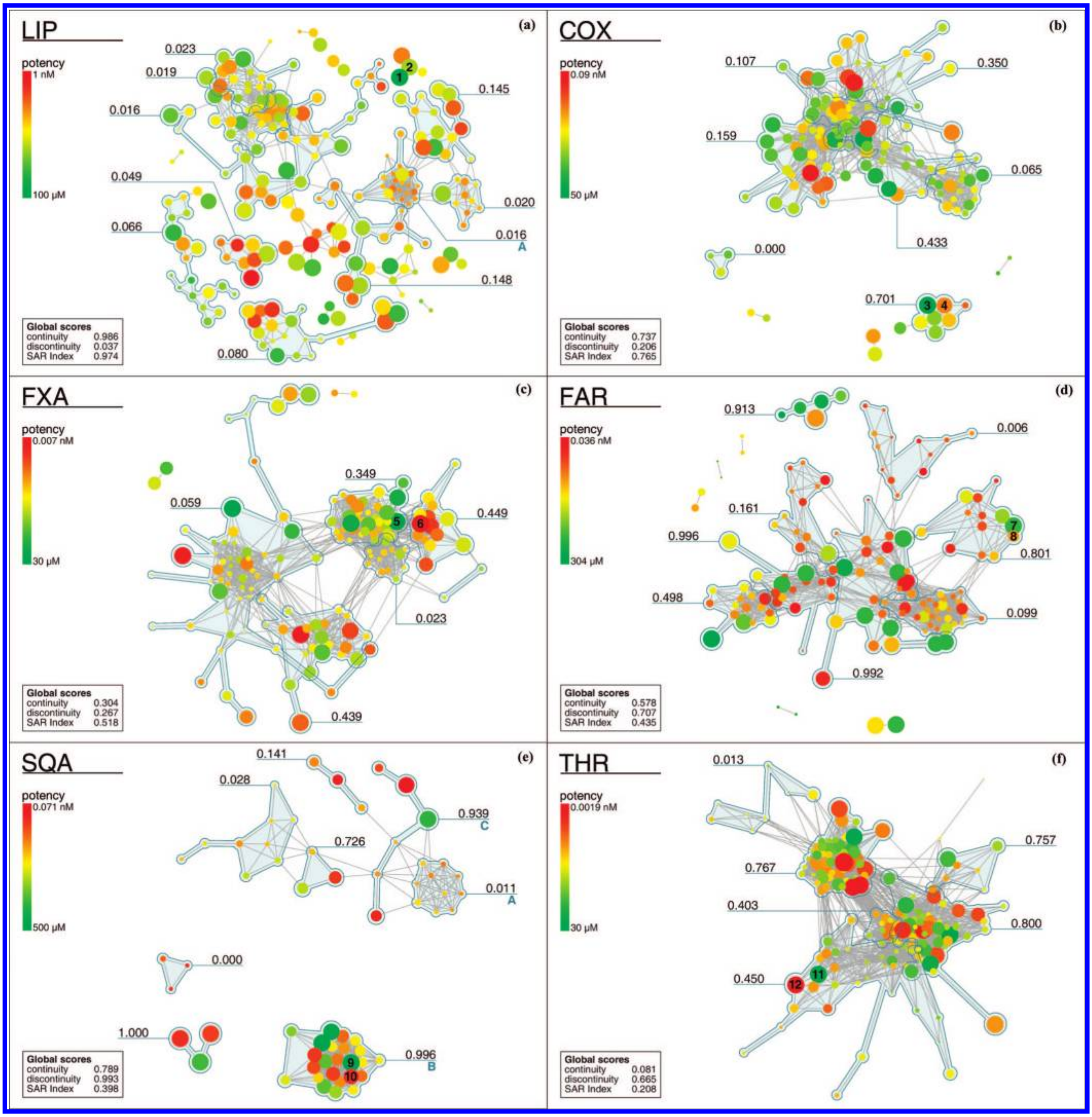

Figure 1. Similarity graphs for six classes of enzyme inhibitors. Nodes represent compounds, and edges are drawn between them if pairwise MACCS Tc values are >0.65. Nodes are color-coded according to potency using a continuous spectrum from green (lowest potency) to red (highest potency) and scaled according to their local compound discontinuity scores (see text for details). Groups of compounds on separate gray "islands" correspond to clusters that are annotated with cluster discontinuity scores. Selected clusters are labeled with capital letters (A, B, or C). Numerically labeled nodes (1, 2, 3,...) correspond to key compounds with high individual discontinuity score, shown in Figure 5. Similarity graphs are shown for the following enzyme inhibitor sets: (a) lipoxygenase (LIP), (b) cyclooxygenase-2 (COX), (c) coagulation factor Xa (FXA), (d) protein farnesyltransferase (FAR), (e) squalene synthase (SQA), (f) thrombin (THR).

highly similar compound pairs. Furthermore, we introduce a cutoff for the potency difference between compound pairs. We consider only compound pairs with potency difference greater than 1 order of magnitude because compounds with very similar potency make only small contributions to the global SAR discontinuity within a compound set.

$$\mathrm{raw}_{\mathrm{disc}} = \frac{\sum_{\{i,j|\mathrm{sim}(i,j)>0.65,\mathrm{potdiff}(i,j)>1,i>j\}} \mathrm{potdiff}(i,j)\,\mathrm{sim}(i,j)}{|\{i,j|\mathrm{sim}(i,j)>0.65,\mathrm{potdiff}(i,j)>1,i>j\}|}$$

The raw scores are converted to Z-scores using the sample mean (μ) and standard deviation (σ) of the scores of the activity class reference panel.

$$\mathrm{zscore}_{\mathrm{cont}} = \frac{\mathrm{raw}_{\mathrm{cont}} - \mu_{\mathrm{cont}}}{\sigma_{\mathrm{cont}}}, \quad \mathrm{zscore}_{\mathrm{disc}} = \frac{\mathrm{raw}_{\mathrm{disc}} - \mu_{\mathrm{disc}}}{\sigma_{\mathrm{disc}}}$$

The Z-scores are then mapped onto the value range [0,1] by calculating the cumulative probability for each score under the assumption of a normal distribution, which yields the final continuity, discontinuity, and SARI scores.

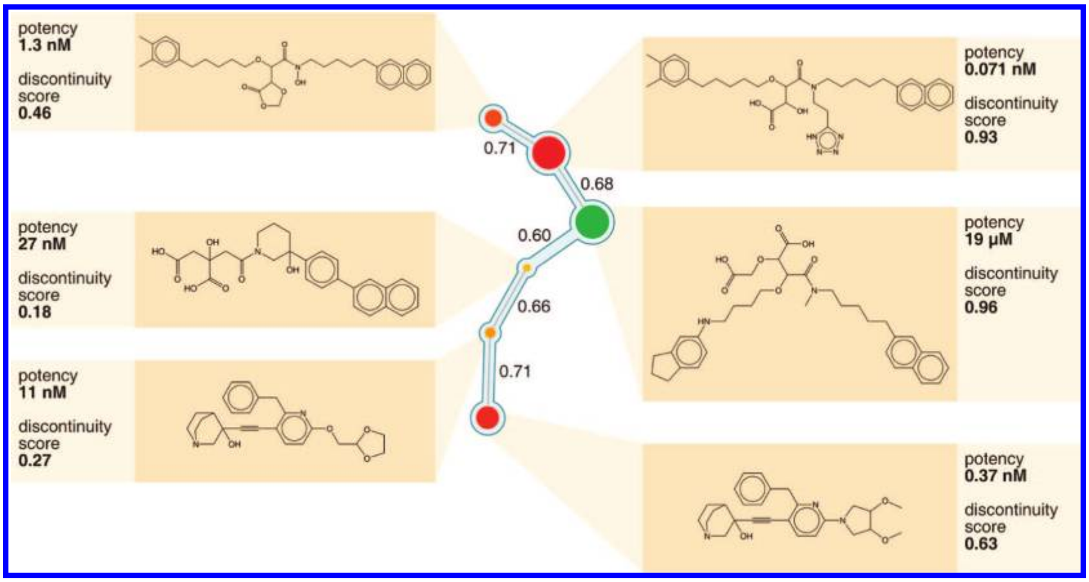

Figure 2. Exemplary NSG compound cluster. Shown are the compounds belonging to cluster C in Figure 1e with their potency values and individual discontinuity scores. Pairwise MACCS Tc values are reported along the edges in the cluster.

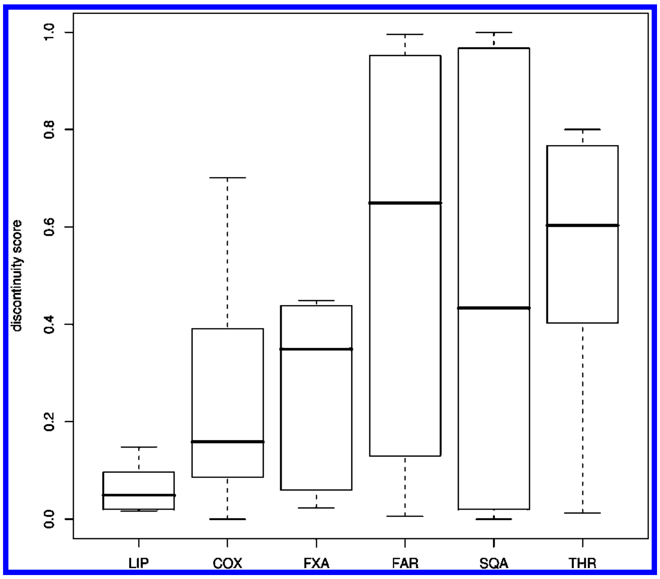

Figure 3. Distribution of cluster discontinuity scores for different activity classes. For each activity class, the distribution of discontinuity scores for all clusters is presented as a box plot. For value distributions, box plots report the smallest observation, lower quartile, median (thick black horizontal line), upper quartile, and largest observation. Overall lowest subset discontinuity scores are present in continuous or heterogeneous-constrained classes. Heterogeneous-relaxed classes show the largest spread of discontinuity scores.

$$\mathrm{score_{cont}} = \Phi(\mathrm{zscore_{cont}}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\mathrm{zscore_{cont}}} \exp\left(-\frac{1}{2}x^2\right)\mathrm{d}x$$

$$\mathrm{score_{disc}} = \Phi(\mathrm{zscore_{disc}}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\mathrm{zscore_{disc}}} \exp\left(-\frac{1}{2}x^2\right)\mathrm{d}x$$

$$\mathrm{SARI} = \frac{1}{2}(\mathrm{score_{cont}} + (1 - \mathrm{score_{disc}}))$$

Local Discontinuity Scores. In order to identify compounds that are responsible for introducing activity cliffs in an activity landscape, we introduce a local variant of the discontinuity score that is calculated on a per compound basis and measures individual contributions to SAR characteristics. The local score of a given active molecule takes into account all compound pairs formed by the molecule and all others within the activity class. The calculation of the discontinuity score for an individual compound also considers only pairs of molecules with a MACCS Tc greater than 0.65. However, in contrast to global SARI calculations, a potency difference of more than 1 order of magnitude is not required here because for the detection of activity cliffs, all potency differences among similar compounds must be taken into account. The local discontinuity score is defined as

$$\mathrm{raw_{disc}}(i) = \frac{\sum_{\{j|\mathrm{sim}(i,j)>0.65, i\neq j\}} \mathrm{potdiff}(i,j)\,\mathrm{sim}(i,j)}{|\{j|\mathrm{sim}(i,j) > 0.65, i \neq j\}|}$$

A local continuity score can be derived in an analogous manner but is not required for the analysis reported herein. Local discontinuity scores are standardized with respect to all compounds within the activity class. Normalization by calculating the cumulative probability on a normal distribution yields the final scores falling into the range [0,1].

Network-like Similarity Graphs. Similarity and potency relationships within an activity class are represented using NSGs. In these graphs, compounds are visualized as nodes, and edges between them display similarity relationships. In order to generate NSGs for sets of active compounds, the MACCS Tc for every compound pair in a class is calculated and the resulting similarity matrix is transformed into an adjacency matrix using a cutoff value of 0.65. If the similarity value of two compounds exceeds the threshold, the corresponding entry in the adjacency matrix is set to 1 and an edge is drawn between the nodes. Otherwise, the entry is set to 0 and the nodes remain disconnected in the graph. Nodes are color-coded by the potency value ($\mathrm{pIC_{50}}$ or $\mathrm{p}K_\mathrm{i}$) of the compounds they represent using a continuous spectrum from green (via yellow) to red, with green indicating lowest potency and red highest potency within a class. The size of each node is scaled according to the local discontinuity score of the compound (i.e., the higher the score, the larger the node). Since the local discontinuity score is normalized with respect to the value distribution of each individual class, both color and size of the nodes reflect the spectrum of values within the given class. Hence, the scale for color and size varies for different classes.

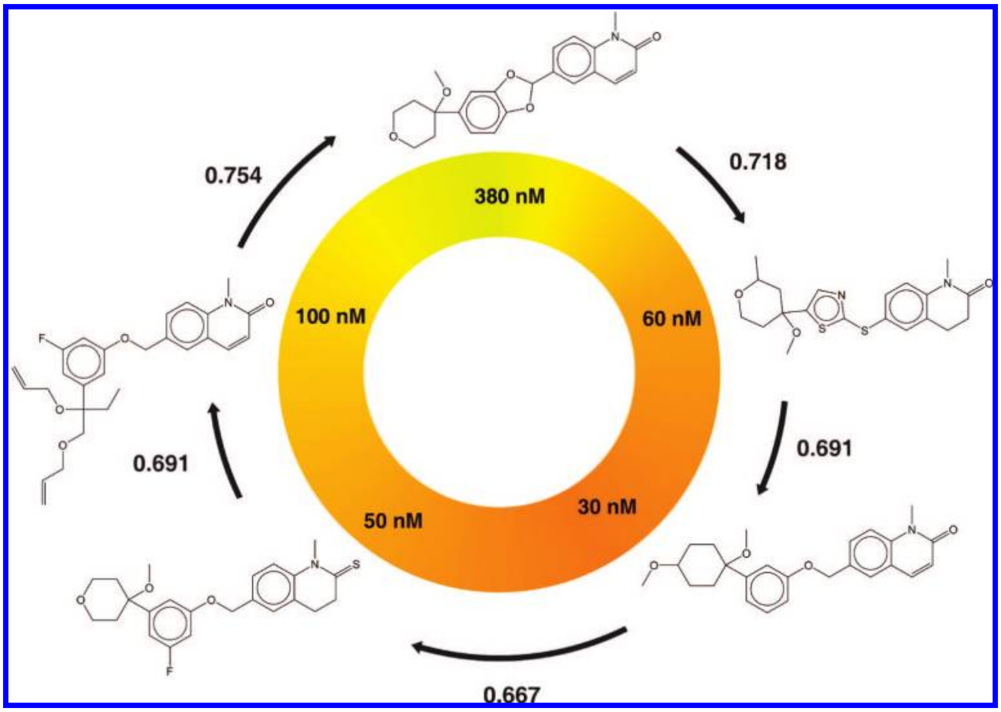

Figure 4. SAR representative compounds. Shown are compounds from cluster A in class LIP representing continuous local SARs over a relatively narrow potency range.

In NSGs, groups of molecules obtained by similarity clustering are highlighted and annotated with the discontinuity score calculated for the set of cluster members. We cluster the molecules of an activity class by hierarchical agglomerative clustering using Ward's minimum variance linkage method,[17] which yields intuitive cluster distributions for our data sets and MACCS Tc values. For this purpose, we transform MACCS Tanimoto similarity into Soergel distance values[16] by subtraction from 1 and use the resulting dissimilarity matrix for clustering. The corresponding dendrograms are pruned at heights between 1 and 2 for different classes. Accordingly, compound pairs within a cluster might have a lower Tc than 0.65, which means that not all nodes within a cluster are necessarily connected by edges. This design introduces an additional level of information that complements the information conveyed by edge connectivity. It enables the detection of remotely related compound subsets and similarity relationships between them. The graphs are generated using the igraph package of R[18,19] utilizing the Fruchterman–Reingold algorithm[20] that computes the distribution of nodes in a two-dimensional plane on the basis of connectivity. Therefore, distances between two nodes are not scaled by similarity values. In the graphical representations, singletons are omitted for clarity.

For a given activity class, network representations can easily be recalculated when more active compounds become available. For all compounds, MACCS fingerprints were calculated using the software package MOE.[21] SARI scores and NSG representations were calculated and visualized using the R software environment.

Activity Classes. Six classes of specific enzyme inhibitors and potency values were assembled from the MDDR that represented different global SAR types on the basis of the SARI classification scheme, as reported in Table 1. The selected classes cover a broad spectrum ranging from globally discontinuous (i.e., low SARI scores) to heterogeneous (intermediate scores) and continuous (high scores) SARs.

Results and Discussion

Our analysis focuses on how to identify SAR features that coexist in activity classes and explore potential relationships between them at the level of compound subsets as well as individual molecules. These studies are ultimately aimed at the identification of key compounds that are major determinants of SAR features and that can be utilized to better understand optimization pathways. In order to characterize subsets of compound activity classes, they are subjected to hierarchical clustering on the basis of pairwise similarity values. Local SARI scoring and NSG representations are developed in order to reveal and describe multiple SAR characteristics within compound classes.

Local SARI. In order to quantitatively describe SAR features and estimate the contributions of individual molecules, we have generated a modified SARI discontinuity score. Local continuity score calculations were not required for our analysis because we primarily focused on compounds introducing activity cliffs. The modified discontinuity score is calculated on a per compound basis and assigns high local scores to molecules that significantly contribute to the global discontinuity in a given compound set. This has made it possible to identify key compounds that largely determine the shape of a given activity landscape, as discussed in detail below.

NSGs and Their Interpretation. As illustrated in Figure 1, NSGs are designed to convey five levels of information. (1) The first level is pairwise similarity relationships. Compounds (nodes) are connected by an edge if their pairwise similarity exceeds a predefined threshold value (MACCS Tc > 0.65). (2) The second level is compound clusters. Groups of compounds that are similar on the basis of hierarchical clustering are displayed on gray "islands". The information provided by edges and islands is complementary; i.e., edges may exist between

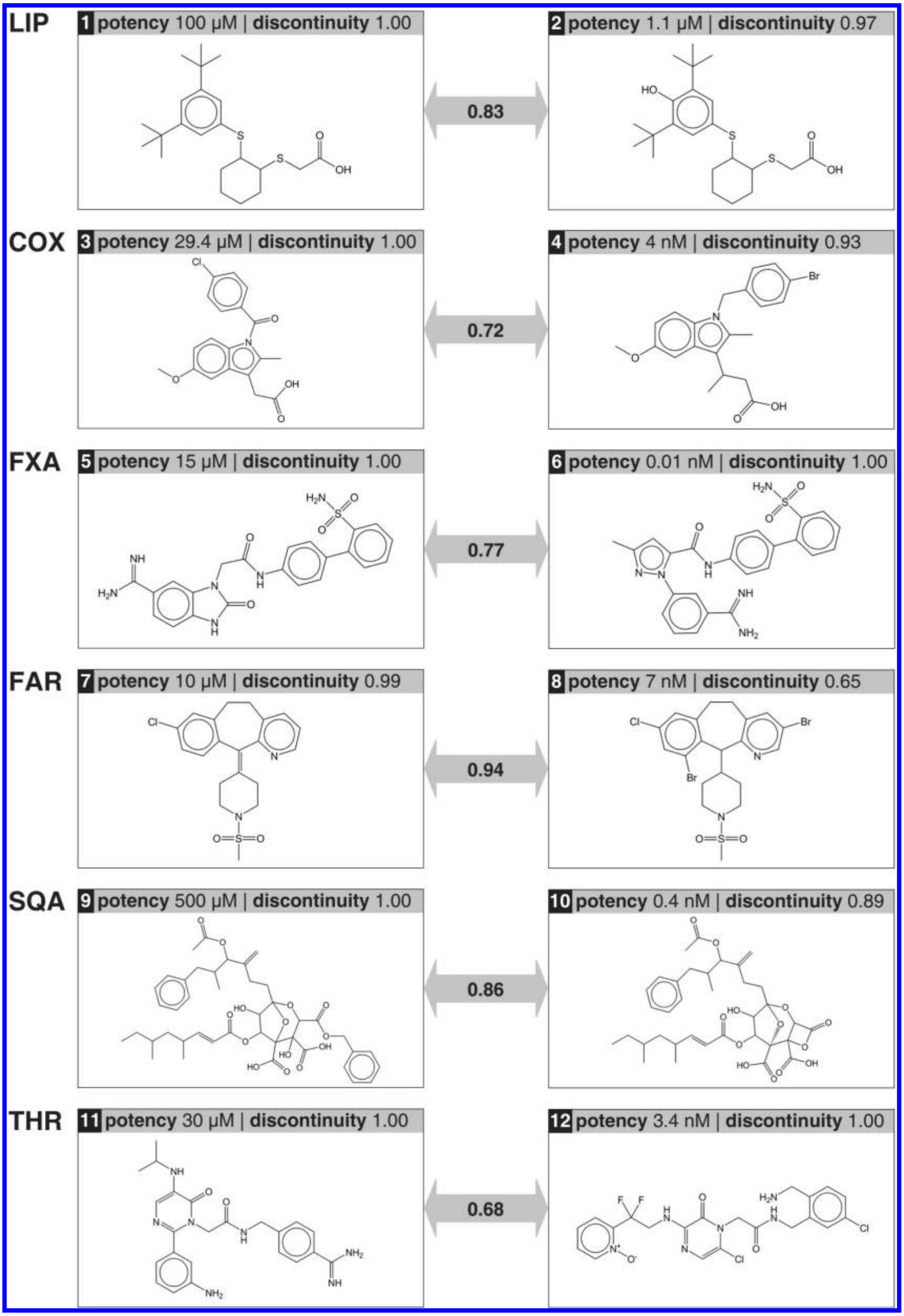
LIP
1 potency 100 μM | discontinuity 1.00
2 potency 1.1 μM | discontinuity 0.97
0.83
COX
3 potency 29.4 μM | discontinuity 1.00
4 potency 4 nM | discontinuity 0.93
0.72
FXA
5 potency 15 μM | discontinuity 1.00
6 potency 0.01 nM | discontinuity 1.00
0.77
FAR
7 potency 10 μM | discontinuity 0.99
8 potency 7 nM | discontinuity 0.65
0.94
SQA
9 potency 500 μM | discontinuity 1.00
10 potency 0.4 nM | discontinuity 0.89
0.86
THR
11 potency 30 μM | discontinuity 1.00
12 potency 3.4 nM | discontinuity 1.00
0.68

Figure 5. Selected key compounds for individual activity classes. Shown are compounds with high local discontinuity score together with another high-scoring neighbor. Potency values and compound discontinuity scores are reported for each molecule as well as MACCS Tc values for compound comparison. Compound numbers correspond to node labels in Figure 1.

compounds belonging to different clusters, but compounds within the same cluster are not always connected by edges. (3) The third level is potency distribution. Color-coding is used to represent the spectrum of potency values within a compound set, ranging from green for lowest potency to red for highest potency. (4) The fourth level is SAR discontinuity and activity cliffs. Nodes are scaled in size according to individual compound contributions to the overall discontinuity within a compound set. Compounds with large nodes are responsible for the introduction of activity cliffs. (5) The fifth level is cluster

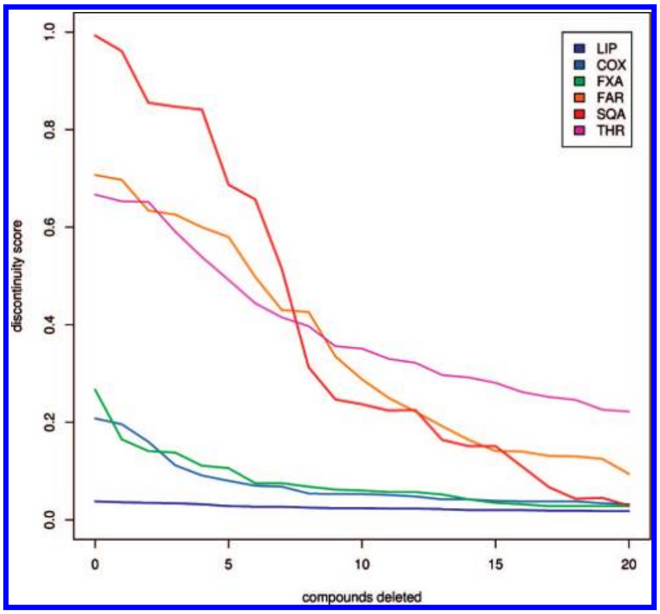

Figure 6. Changing SAR characteristics after removal of key compounds. For each activity class, compounds with highest local discontinuity scores were iteratively removed from the compound sets and global discontinuity scores were recalculated following each iteration.

contributions to SAR characteristics. Compound clusters are annotated with their cluster discontinuity score, which makes it possible to highlight different SAR regions within a compound set.

Topology of NSGs. The topology of NSGs is determined by pairwise similarity relationships and their distribution within activity classes. Figure 1 shows that different activity classes produce NSGs of different topology and that distinct topologies are also observed for compound sets having similar global SAR characteristics. For example, the LIP and COX compound sets studied here display globally continuous SARs, as indicated by high SARI scores. LIP is characterized by a high degree of intraclass structural diversity (average pairwise MACCS Tc = 0.36). Its NSG consists of distinct subgraphs, and many compounds are only sparsely connected, but several densely connected clusters with very low discontinuity scores are observed (Figure 1a). Thus, intraclass structural diversity in part results from chemically different series of active compounds. Compared to LIP, COX is less structurally diverse (average MACCS Tc = 0.45) and its NSG is overall more densely connected (Figure 1b). It contains a major network component and only a few peripheral clusters and compound pairs.

In heterogeneous-constrained SAR types, structural variations of active compounds occur within the boundaries of moderate activity cliffs (i.e., functional groups involved in key interactions are conserved, while other molecular regions are more variable). Class FXA belongs to this category (characterized by low global continuity and discontinuity scores and an intermediate SARI score; see Table 1), and its NSG exhibits a few densely connected components that are organized around apparent activity cliffs (Figure 1c).

In heterogeneous-relaxed SAR types, continuous and discontinuous SAR components coexist.[8] This SAR type is characterized by high global continuity and discontinuity scores and an intermediate SARI score. FAR provides a representative example. The structural diversity within this class is comparable to COX and so is its NSG topology (Figure 1d). By contrast, SQA has intraclass structural diversity comparable to FAR and also belongs to the heterogeneous-relaxed SAR category but produces an NSG of different topology (Figure 1e). The graph consists of several distinct components that are clearly separated from each other. These subgraphs are well-defined and correspond to structurally distinct subsets of compounds that display different SAR characteristics, as indicated by their cluster discontinuity scores. However, although FAR and SQA display different NSG topologies, Figure 1d and Figure 1e also show that in both cases individual compounds are found to form edges between distinct subgraphs with very different discontinuity scores. Thus, structurally similar compounds can be identified in NSGs that are involved in different SAR relationships and represent "chemical bridges" between SAR islands. Furthermore, in Figure 1e, clusters can be seen that are prototypic instances of flat SARs (cluster A, collection of small nodes at intermediate potency levels) or pronounced activity cliffs (cluster B, large nodes covering wide potency ranges).

The set of THR inhibitors used in this study represents the structurally most homogeneous class (average MACCS Tc = 0.55) but the most discontinuous global SAR (characterized by a low SARI score). Its NSG contains a single and densely connected major network component (and only one peripheral cluster). Consistent with its SAR type, many individual compounds are found to make large local contributions to SAR discontinuity and form activity cliffs.

Irrespective of their topology, all NSGs show a clear correspondence between compound clusters and graph or subgraph communities. Thus, clusters can serve as a basis for studying local SAR characteristics.

Cluster SARs. In order to analyze and compare compound subset-dependent SAR features, SARI discontinuity scores were calculated for clusters to quantify their differences, as shown in Figure 1. Individual clusters can be isolated from NSGs and analyzed separately. This makes it possible to select compound subsets on the basis of their SAR characteristics and study their composition in detail. Figure 2 shows an example for class SQA. This cluster has a high discontinuity score and should therefore include structurally related compounds with distinct potency differences. For a detailed analysis of this cluster, similarity relationships are displayed and the local discontinuity scores are reported for each compound. Consistent with our expectation, structurally related representatives of two compound series can be easily identified. The cluster contains two compounds that make large local contributions to discontinuity (corresponding to the large green and red nodes) and are largely responsible for its overall discontinuous SAR behavior. These compounds can be considered to mark the beginning (green) and end (red) of a lead optimization pathway. As can be seen in Figure 1e, this cluster also forms chemical bridges to other clusters of inhibitors with continuous or discontinuous SAR characteristics. Thus, analysis of cluster SARs and intercluster similarity relationships makes it possible to identify compounds that represent different local SARs within activity classes and delineate alternative optimization pathways.

Cluster SARs versus Global SARs. The analysis of local SAR discontinuity is essential to identify activity cliffs and key compounds that are responsible for them. In order to relate discontinuity within individual compound subsets to global SAR features, we have calculated the distribution of cluster discontinuity scores for each activity class, as shown in Figure 3. Classes representing different SAR types were found to display characteristic score distributions. Compound classes having globally continuous or heterogeneous-constrained SARs lack steep activity cliffs. Accordingly, the cluster discontinuity scores

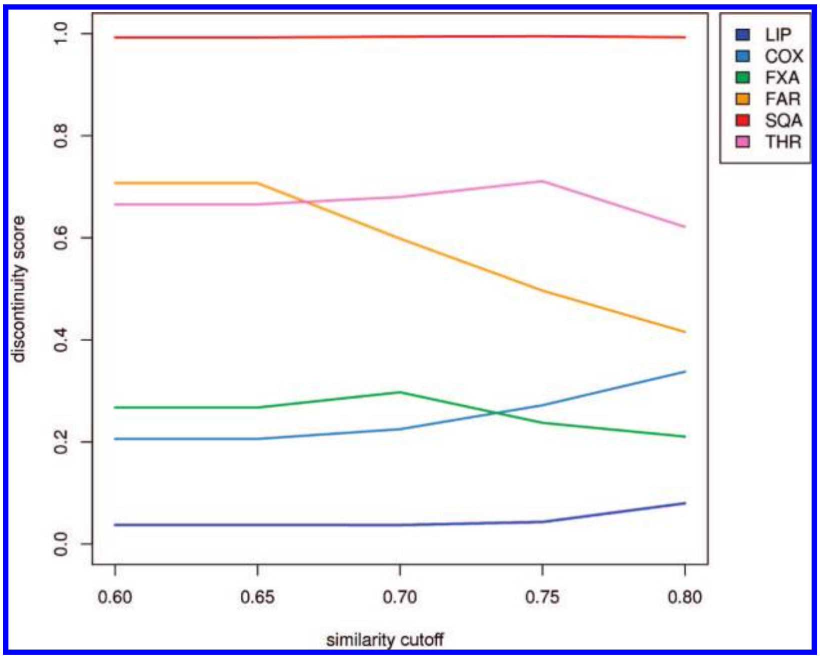

Figure 7. Global discontinuity scores for increasing similarity cutoff levels. For each activity class, global discontinuity scores are shown for similarity cutoff levels increasing from 0.6 to 0.8.

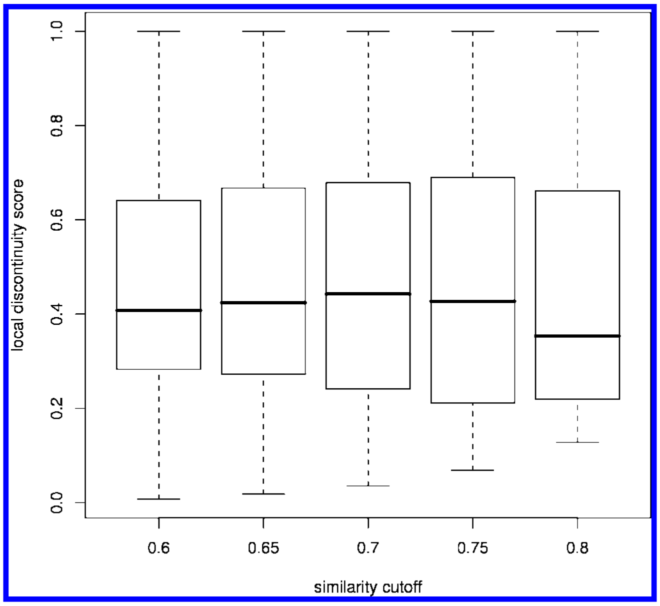

Figure 8. Distributions of local discontinuity scores for different similarity cutoff levels. Local discontinuity scores were calculated for each compound of the activity classes studied here using increasing similarity cutoff values. Score distributions for different cutoff levels are presented as box plots.

for these classes range from low to intermediate values. For the continuous class LIP, several compound clusters are assigned discontinuity scores at the lower end of the spectrum. For example, cluster A in Figure 1a obtains a low score of 0.016. Accordingly, this cluster contains highly potent compounds with little potency variation (Figure 4), which results in overall low discontinuity in the cluster SARs.

In globally discontinuous classes like THR, clusters show significantly higher discontinuity scores, as expected (Figure 3). A few clusters having very high scores form strong activity cliffs that dominate global SAR features. Cluster scores in heterogeneous-relaxed compound classes show the largest variations, due to the coexistence of continuous and discontinuous SAR components. The NSG of class SQA in Figure 1e nicely illustrates this heterogeneous SAR character and contains clusters that correspond to continuous or even flat SARs (cluster A) or, by contrast, represent prime examples of a rugged activity landscape that contains similar compounds with large potency spread, corresponding to steep activity cliffs (B).

Local SAR Discontinuity and Key Compounds. In order to focus the analysis on individual activity cliffs, discontinuity scores were calculated on a per compound basis. For all activity classes, highly and weakly potent key compounds making large contributions to discontinuity were identified, irrespective of the global SAR phenotype. Selected key compounds are labeled in Figure 1 and are shown in Figure 5. These compounds are similar but have dramatic differences in potency and are thus activity cliff markers that are easily identified in NSGs. Because local scores are normalized relative to each individual compound class, key compounds from different classes typically represent different levels of discontinuity, depending on the overall score distribution within the classes. The ability to identify activity cliff markers for all activity classes studied here underlines the value of local SARI discontinuity scores. In order to estimate their influence on global SARs, compounds with highest local scores were selected and iteratively removed from activity classes and global and local SARI scores were recalculated after each step. The results in Figure 6 show that for discontinuous and heterogeneous-relaxed SAR types the global discontinuity score constantly decreases and thus shifts global SARs more toward the continuous range. Thus, in these cases, key compounds with high local discontinuity score also strongly influence global SAR characteristics. By contrast, for continuous or heterogeneous-constrained classes, moderate activity cliffs identified on the basis of local scores only have limited influence on global SAR characteristics.

Control Calculations. In order to assess the sensitivity of the results to chosen parameter settings, we carried out a number of control calculations. First, we investigated the influence of the similarity cutoff on global and local discontinuity scores. Therefore, the Tc cutoff value was varied from 0.6 to 0.8 in steps of 0.05. Figure 7 shows that global discontinuity scores

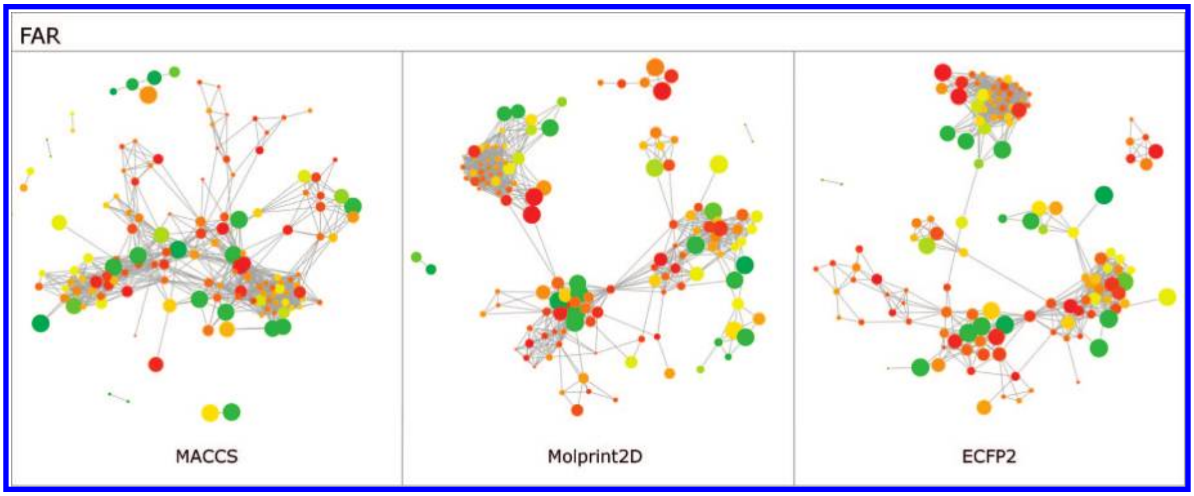

Figure 9. Similarity graphs using different fingerprints. As representative examples, similarity graphs for class FAR are shown calculated using the MACCS, Molprint2D, and ECFP2 fingerprints.

remained largely stable over this value range, with the exception of class FAR where the score decreased notably for cutoff values above 0.65. This behavior is due to the fact that FAR contains many compounds with pairwise similarity close to 0.65 that form activity cliffs that are not accounted for at higher threshold levels.

Local discontinuity scores were more sensitive to variations in the similarity cutoff. For some compounds, scores were found to vary substantially yielding standard deviations of up to 0.49. On average, however, standard deviations of local scores only amounted to 0.13 for varying similarity threshold levels. Moreover, as illustrated in Figure 8, the overall distribution of local discontinuity scores for all classes remained essentially unaffected.

Second, considering the fact that SAR descriptions generally depend on the nature of the chosen molecular representation, we also compared the MACCS-based results with two different molecular fingerprints as similarity measures: Molprint2D[22,23] and ECFP2.[24] Similarity cutoffs were adjusted to the similarity distribution for the individual fingerprints. Representative similarity graphs for all three fingerprint types are shown in Figure 9. We observed that despite differences in network connectivity, the global topological characteristics were preserved for the compound classes studied here. For global and local SARI scores, a correlation was observed for different fingerprint representations (average pairwise correlation of global SARI scores, 0.73; average pairwise correlation of local discontinuity scores, 0.68).

Conclusions

A systematic study of local SARs has been carried out by comparative analysis of potency distributions and similarity relationships among different classes of active compounds. The network-like similarity graphs and local SAR indices introduced herein make it possible to dissect SAR phenotypes and relate local and global SAR features to each other, both in qualitative and quantitative terms. In NSGs, compound subsets often form separate communities that are related by different local SARs. However, communities with different SAR character are frequently connected by chemical bridges, i.e., chemically similar molecules that occur in different SAR contexts. Our NSG-SARI analysis also reveals that a few key compounds are in some instances capable of shaping the activity landscape of a collection of active molecules. This makes it possible to identify compounds that are activity cliff markers and often represent the start and end points of optimization efforts. The identification of such compounds and the opportunity to capture SAR characteristics at different levels render the NSG-SARI framework attractive for practical applications in medicinal chemistry. It is readily possible to elucidate multiple SAR components present in large data sets and prioritize compound subsets for further analysis and chemical optimization.

References

(1) Kubinyi, H. Similarity and Dissimilarity. A Medicinal Chemist's View. *Perspect. Drug Discovery Des.* **1998**, *9*−*11*, 225–252.

(2) Maggiora, G. M. On Outliers and Activity Cliffs. Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535–1535.

(3) Eckert, H.; Bajorath, J. Molecular Similarity Analysis in Virtual Screening: Foundations, Limitations, and Novel Approaches. *Drug Discovery Today* **2007**, *12*, 225–233.

(4) Johnson, M., Maggiora, G. M., Eds. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, 1990.

(5) Shanmugasundaram, V.; Maggiora, G. M. Characterizing Property and Activity Landscapes Using an Information-Theoretic Approach. Presented at the 222nd American Chemical Society National Meeting, Division of Chemical Information, 2001; Abstract No. 77.

(6) 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.

(7) Peltason, L.; Bajorath, J. Molecular Similarity Analysis Uncovers Heterogeneous Structure−Activity Relationships and Variable Activity Landscapes. *Chem. Biol.* **2007**, *14*, 489–497.

(8) Peltason, L.; Bajorath, J. SAR Index: Quantifying the Nature of Structure−Activity Relationships. *J. Med. Chem.* **2007**, *50*, 5571–5578.

(9) Guha, R.; Van Drie, J. H. Structure−Activtiy Landscape Index: Identifying and Quantifying Activtiy Cliffs. *J. Chem. Inf. Model.* **2008**, *48*, 646–658.

(10) Mestres, J.; Martin-Couce, L.; Gregori-Puigjane, E.; Cases, M.; Boyer, S. Ligand-Based Approaches to in Silico Pharmacology: Nuclear Receptor Profiling. *J. Chem. Inf. Model.* **2006**, *46*, 2725–2736.

(11) Paolini, G. V.; Shapland, R. H. B.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global Mapping of Pharmacological Space. *Nat. Biotechnol.* **2006**, *24*, 805–815.

(12) Hert, J.; Keiser, M. J.; Irwin, J. J.; Oprea, T.; Shoichet, B. K. Quantifying the Relationships among Drug Classes. *J. Chem. Inf. Model.* **2008**, *48*, 755–765.

(13) Yildirim, M. A.; Goh, K.-I.; Cusick, M. E.; Barabási, A.-L.; Vidal, M. Drug−Target Network. *Nat. Biotechnol.* **2007**, *25*, 1119–1126.

(14) *Molecular Drug Data Report (MDDR)*, version 2005.2; Symyx Software: San Ramon, CA, 2005.

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

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

(17) Ward, J. H. Hierarchical Grouping To Optimize an Objective Function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244.

(18) R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2008.

(19) Csardi, G. *igraph library*, version 0.5; Budapest, Hungary, 2008; http://cneurocvs.rmki.kfki.hu/igraph (accessed March 7, 2008).

(20) Fruchterman, T. M. J.; Reingold, E. M. Graph Drawing by Force-Directed Placement. *Software—Pract. Experience* **1991**, *21*, 1129–1164.

(21) *Molecular Operating Environment (MOE)*, version 2007.09; Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2007.

(22) 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.

(23) MOLPRINT 2D. http://www.molprint.com (accessed June 2006).

(24) *SciTegic Pipeline Pilot Student Edition*, version 6.1.5; Accelrys, Inc.: San Diego, CA, 2007.

JM800867G