CHEMMEDCHEM

DOI: 10.1002/cmdc.200900300

From Structure–Activity to Structure–Selectivity Relationships: Quantitative Assessment, Selectivity Cliffs, and Key Compounds

Lisa Peltason, Ye Hu, and Jürgen Bajorath*[a]

The exploration of structure–activity relationships (SARs) in chemical lead optimization is mostly focused on activity against single targets. Because many active compounds have the potential to act against multiple targets, achieving a sufficient degree of target selectivity often becomes a major issue during optimization. Herein we report a data analysis approach to explore compound selectivity in a systematic and quantitative manner. Sets of compounds that are active against multiple targets provide a basis for exploring structure–selectivity relationships (SSRs). Compound similarity and selectivity data are analyzed with the aid of network-like similarity graphs (NSGs), which organize molecular networks on the basis of similarity relationships and SAR index (SARI) values. For this purpose, the SARI framework has been adapted to quantify SSRs. Using sets of compounds with differential activity against four cathepsin thiol proteases, we show that SSRs can be quantitatively described and categorized. Furthermore, local SSR environments are identified, the analysis of which provides insight into compound selectivity determinants at the molecular level. These environments often contain "selectivity cliffs" formed by pairs or groups of similar compounds with significantly different selectivity. Moreover, key compounds are identified that determine characteristic features of single-target SARs and dual-target SSRs. The comparison of compounds involved in the formation of selectivity cliffs often reveals chemical modifications that render compounds target selective.

Introduction

Structure–activity relationships of small molecules are generally dependent on the compound class and target, and are often highly complex.[1–3] Their individual nature and variability makes it difficult to describe and analyze SARs in a consistent and quantitative manner.[3] Ultimately, SAR analysis aims at predicting potent compounds, which explains the popularity of quantitative SAR analysis methods in medicinal chemistry.[4,5] However, compound potency represents only one of several measures of successful lead optimization, including metabolic stability, non-toxicity, and selectivity. Traditionally, target selectivity of lead compounds has been considered a stringent requirement for their ability to become drug candidates.[6] However, increasing evidence suggests that the majority of drugs and other biologically active compounds are likely to act on more than one target, and often many.[6–9] Such insights are beginning to change our view of compound selectivity and its importance for therapeutic intervention. In fact, exclusive target selectivity resulting from "all or nothing" binding events is probably an exception rather than the rule for biologically active compounds. Nevertheless, achieving single-target compound selectivity continues to be highly valuable in drug discovery, for example, in the search for molecules that are active against microbial target proteins for which orthologous proteins might exist. However, in many instances, apparent selective inhibition or antagonism is likely to result from differences in compound potency against multiple targets. This is particularly relevant for closely related members of protein families in which a series of active compounds might produce multi-target SARs.[10] Such multi-target SARs ultimately determine various degrees of compound selectivity.

Large experimental efforts are required to test active compounds against arrays of targets and determine selectivity profiles.[10] Thus, it is not surprising that there has been increasing interest in computational approaches to analyze and predict compound selectivity.[11] However, at present the computational study of selectivity is still in its infancy.[11] For SAR analysis, computational methods have recently been introduced that make it possible to profile and compare SARs on a large scale and identify key compounds that determine SAR features.[12,13] These approaches conceptually differ from other computational methods that attempt to fit compound data sets to linear or nonlinear models of SARs.[14] Rather, SAR analysis functions such as the SAR index (SARI)[12] attempt to extract available SAR information directly from compound data sets. Hence, such methods represent a data-oriented analysis strategy.[14] Similarity and potency relationships between active compounds can be displayed in molecular network representations that make it possible to graphically access SAR features, compare SARs, and identify SAR determinants.[15]

[a] L. Peltason,[+] Y. Hu,[+] Prof. Dr. J. Bajorath
Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität Bonn, Dahlmannstr. 2, 53113 Bonn (Germany)
Fax: (+49) 228-2699-341
E-mail: bajorath@bit.uni-bonn.de

[+] These authors contributed equally to this work.

Applying similar data-oriented analysis methods to the study of molecular selectivity is considered an attractive goal because compound selectivity can in many instances be rationalized as a result of multi-target SARs. Taking into account that there are currently no computational methods available to systematically analyze and compare compound selectivity profiles, much information about selectivity determinants might be gained through data mining. Herein we report a first step in this direction by adapting and extending numerical and graphical SAR analysis functions for the study of structure–selectivity relationships. We report that selectivity relationships within compound data sets can be categorized in analogy to SARs and graphically compared. In compound data sets with experimental measurements against four cathepsins, different global and local SSRs emerge, and individual compounds are identified that make large contributions to single-target SARs and dual-target SSRs. Such key compounds can be readily selected for further chemical exploration. In addition, the analysis of local SSR environments makes it possible to identify structural modifications that are selectivity determinants. Hence, a major focal point of systematic SSR analysis is to aid in compound selection and analogue design.

Materials and Methods

Selectivity data sets

For the comparative study of SARs and SSRs, we have analyzed inhibitor sets for cathepsin (cat) B, K, L, and S for which potency measurements against at least two to four proteases were available. These cathepsin inhibitor data were taken from previously reported publicly available compound collections assembled from original literature sources for chemical biology applications.[16] A pool of 312 inhibitors was subdivided into four partly overlapping sets of compounds focusing on four different target pairs: cat L–cat B (LB), SB, KL, and SK. For 97 inhibitors, potency values were available for all four cathepsins, and these compounds thus occurred in all four subsets. Table 1 summarizes the composition of our compound sets.

Compound selectivity was determined on the basis of differential potency against target pairs. Selectivity values of compounds selective for target A over target B were calculated as the difference between their pK_i or pIC_{50} values:

$$S_i = P_i(A) - P_i(B) \quad (1)$$

for which S_i is the selectivity value of compound i, $P_i(A)$ is the potency value of compound i for target A, and $P_i(B)$ is the potency value of compound i for target B. Compounds with a selectivity value >1.7 were considered selective for target A over B, and compounds with a value of <-1.7 were selective for target B over A. This threshold corresponds to a 50-fold difference in potency. Compounds falling within this range were considered nonselective.

Table 1. Compound sets with potency values for two related targets.[a]

Set Identifier	Target A	Target B	#Compds[b]	Potency Range (A)	Potency Range (B)	Selectivity Range (A/B)
LB	cat L	cat B	159/26/4	3.82–10.40	3.00–8.07	−2.21–3.17
SB	cat S	cat B	142/34/3	3.82–9.73	3.82–8.07	−2.29–4.63
KL	cat K	cat L	234/54/18	4.00–11.05	3.82–10.40	−5.08–4.96
SK	cat S	cat K	248/76/47	3.00–9.89	3.00–11.05	−5.37–4.54

[a] The composition of compound data sets and potency and selectivity value ranges are reported; "cat" stands for cathepsin. [b] #Compds gives the number of compounds per set: the first number reports the total number of compounds, the second the number of compounds selective for target A, and the third the number of compounds selective for target B; the remaining compounds in each set are classified as nonselective.

Molecular similarity assessment

The similarity between two molecules was calculated using the Tanimoto coefficient (Tc)[17] for comparison of their MACCS fingerprint representations.[18] The set of 166 publicly available MACCS structural keys was found to produce chemically meaningful and easily interpretable results in SAR profiling[12] and molecular network analysis,[15] and was thus used throughout this analysis. However, the methods presented herein are applicable in combination with any structural descriptors and similarity measures.

SARI scores

SARI scores for sets of active compounds were calculated as described previously.[12] The SARI function is composed of two individual scores that account for smooth and rough regions within an activity landscape. The continuity score reflects the presence of structurally diverse compounds with similar activity, which corresponds to a "continuous" SAR. It is calculated as the weighted arithmetic mean of pairwise compound dissimilarity within a data set. This weighting scheme emphasizes compound pairs with high overall potency and low difference in potency [Eq. (2) and Eq. (3)].

$$cont = weighted\ \underset{\{i,j|i>j\}}{mean}\left(\frac{1}{1+sim(i,j)}\right) = \frac{\sum_{\{i,j|i>j\}} w_{ij}\frac{1}{1+sim(i,j)}}{\sum_{\{i,j|i>j\}} w_{ij}} \quad (2)$$

$$w_{ij} = \frac{P_i \cdot P_j}{1+\left|P_i - P_j\right|} \quad (3)$$

Here, sim(i,j) denotes the MACCS Tc similarity between compounds i and j, and P_i gives the potency value of compound i (as pK_i or pIC_{50} value).

The discontinuity score quantifies the occurrence of activity cliffs that give rise to "discontinuous" SARs by calculating the average potency difference between compound pairs that are structurally similar with respect to a predefined threshold, here set to a MACCS Tc value of 0.65. Activity cliffs within an activity

landscape are formed by structurally similar compounds having large differences in potency. Therefore, in discontinuity score calculations, potency differences are scaled by pairwise similarity in order to emphasize cliffs formed by highly similar compounds. In addition, only compound pairs with a potency difference of more than one order of magnitude are considered.

$$disc = \underset{\left\{i,j \middle| \begin{smallmatrix} sim(i,j)>0.65, \\ |P_i-P_j|>1, i>j \end{smallmatrix}\right\}}{mean} \left(|P_i - P_j| \cdot sim(i,j)\right) = \frac{\sum_{\left\{i,j \middle| \begin{smallmatrix} sim(i,j)>0.65, \\ |P_i-P_j|>1, i>j \end{smallmatrix}\right\}} |P_i - P_j| \cdot sim(i,j)}{|\{i,j | sim(i,j) > 0.65, |P_i - P_j| > 1, i > j\}|} \quad (4)$$

The "raw" continuity and discontinuity scores are normalized using the score distribution of a reference set of compound classes.[15] Raw scores are first converted into conventional *Z* scores by using the sample mean and standard deviation of the reference score distribution and then mapped onto the value range [0,1] by calculating the cumulative probability for each *Z* score under the assumption of a normal distribution. The final SARI value combines the continuity and discontinuity scores and also adopts values from 0 to 1.

$$SARI = \frac{1}{2}(cont_{norm} + (1 - disc_{norm})) \quad (5)$$

Compound discontinuity scores

In order to assess the contribution of individual compounds to the formation of activity cliffs, we calculate a modified discontinuity score on the basis of each compound and all compounds that are similar to it.[15] For a given compound, we consider all compound pairs formed by this compound and its neighbors (i.e., compounds above a predefined similarity threshold).

$$disc(i) = \underset{\{j|j\neq i, sim(i,j)>0.65\}}{mean} \left(|P_i - P_j| \cdot sim(i,j)\right) = \frac{\sum_{\{j|j\neq i, sim(i,j)>0.65\}} |P_i - P_j| \cdot sim(i,j)}{|\{j|j\neq i, sim(i,j) > 0.65\}|} \quad (6)$$

Here no potency difference threshold is applied because for the assessment of discontinuity contributions of individual compounds, all potency differences among similar compounds must be taken into account. Compound discontinuity scores are also normalized by the calculation of *Z* scores and cumulative probabilities to adopt values between 0 and 1. However, in contrast to the normalization of scores for sets of compounds, the distribution of all compound scores within the data set serves as the reference for score normalization.[15]

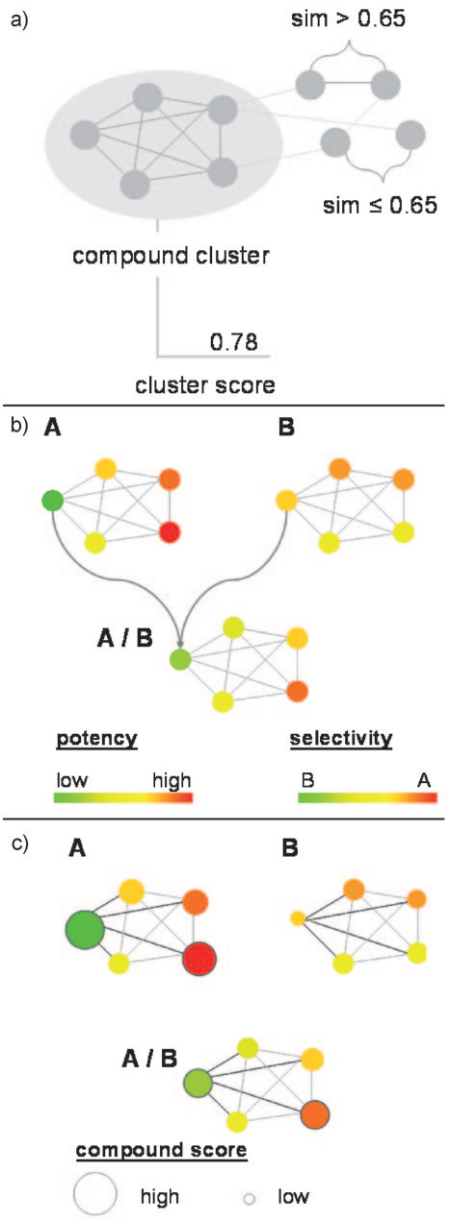

Figure 1. Schematic representation of NSG information levels: a) Nodes represent compounds and are connected if their MACCS Tc similarity exceeds a predefined threshold value. In addition, compounds are clustered based on pairwise similarity values. For compound clusters, SARI discontinuity scores are calculated. b) For a data set with known potency values for two targets A and B, three NSGs are generated using potency values for target A (graph A) or for target B (graph B) and using selectivity values for target A over target B (graph A/B). Node colors in potency NSGs correspond to potency values, and node colors in selectivity NSGs to selectivity values that are derived from the potency differences. c) For each compound in an NSG, a compound discontinuity score is calculated by relating its potency or selectivity value to the corresponding values of similar compounds (indicated by dark edges). Node sizes are then scaled according to the magnitude of compound scores. Pairs of compounds represented by large red and green nodes are key compounds that mark an "activity cliff" in potency NSGs (top) or a "selectivity cliff" in selectivity NSGs (bottom). Key compounds in graphs A and A/B are encircled.

Network-like similarity graphs

NSGs are used to visualize similarity and potency relationships within a compound data set.[15] Figure 1 shows a schematic representation of these molecular graphs and the information they convey. In NSGs, compounds are represented by nodes that are connected by an edge if the similarity between the

corresponding compounds exceeds a predefined similarity threshold (MACCS Tc > 0.65). The topology of NSGs is calculated on the basis of node connectivity using the Fruchterman–Reingold layout algorithm[19] implemented in the R *igraph* package.[20,21] Therefore, distances between two nodes are not scaled by similarity values but rather indicate how densely the nodes within regions of the network are connected by edges. Thus, the network topology implicitly reflects the information conveyed by edges; however, edges are drawn to indicate which nodes correspond to compounds that are similar to each other.

In addition, subsets of similar molecules are obtained by clustering the compounds using their MACCS Tc similarity and Ward's clustering algorithm,[22] and for each compound cluster, SARI scores are calculated (Figure 1 a). SARI scores calculated for compound clusters report local SAR or SSR features and are complementary to the global scores calculated for an entire compound data set. Local SARs or SSRs, that is, SARs or SSRs found in subsets of similar compounds present in defined network regions, often display distinct characteristics and are hereafter referred to as "local SAR/SSR environments". Moreover, compounds in a cluster often form well-defined sub-networks that are densely connected and display distinct topology and SAR/SSR features. These sub-graph structures are also termed "(local) communities". Notably, the applied clustering algorithm might assign compounds to the same cluster even if they are not connected by an edge (because their similarity value is below the threshold), and compounds that are connected by an edge might be assigned to different clusters. Hence, compound clusters complement the binary similarity information provided by edges. Nodes are color-coded according to the potency values of the compounds using a continuous spectrum from green via yellow to red, with green indicating lowest potency and red the highest potency within a set (Figure 1 b). Furthermore, for each compound in the data set, the compound discontinuity score is calculated, and nodes in the NSG are then scaled in size according to compound scores (i.e., the higher the score, the larger the node; see Figure 1 c).

Results and Discussion

Potency and selectivity NSGs

For each data set in Table 1, three different graph representations were generated. First, we separately calculated SARI scores and NSGs by using the potency information for individual targets, thus producing two "potency NSGs": NSG_A and NSG_B (labeled "A" and "B" in Figure 1 b). These potency NSGs were used to characterize single-target SARs. Potency-based compound scores were normalized with reference to all scores calculated using both activities for the given data set. Due to this common normalization scheme, compound scores and node sizes can be directly compared in NSG_A and NSG_B. For this purpose, the same color spectrum was also used for nodes in NSG_A and NSG_B to represent their potency values. Thus, these values ranged from the lowest to the highest potency of a compound active against one or the other target. Then, "selectivity NSGs" were calculated by using the selectivity values for target A over target B, NSG_{AB} ("A/B" in Figure 1 b), and used to explore dual-target SSRs. Nodes are colored according to selectivity values using a spectrum from red for the highest observed selectivity for target A to green for the corresponding inverse selectivity value. Hence, yellow nodes represent nonselective compounds, that is, compounds with similar potency for both targets (Figure 1 b). For selectivity NSGs, SARI scores were calculated by using selectivity values instead of potency values. Thus, the selectivity-based SARI discontinuity scores identify "selectivity cliffs" formed by structurally similar compounds having significantly different selectivity (encircled nodes in Figure 1 c, graph "A/B"). Compound scores in NSG_{AB} were normalized with respect to all selectivity-based scores in the data set and thus cannot be directly compared with scores in potency NSGs (or other selectivity NSGs). Global selectivity-based SARI scores for the entire data set and for individual clusters were normalized relative to the same reference class panel as the potency-based global scores. This normalization procedure was appropriate because selectivity values result from potency differences and, consequently, selectivity-based scores are of the same dimension as their potency-based counterparts.

Global SAR and SSR features

Selectivity-based SARI scores fall into the range between 0 and 1, and are measures of SSR continuity and discontinuity, in analogy to potency-based SARI scores and SARs.[12] High scores close to 1 are indicative of continuous SSRs, where gradual changes in molecular structure are accompanied by moderate changes in compound selectivity, whereas low scores close to 0 reflect discontinuous SSRs, where similar compounds have different selectivity. Intermediate scores around 0.5 are characteristic of heterogeneous SSRs that combine continuous and discontinuous SSR elements.

We first determined the global SAR and SSR categories of the four compound sets. SARI values are reported in Table 2.

Table 2. Global SARI scores for potency and selectivity values.[a]

Set Identifier	Potency (A)			Potency (B)			Selectivity (A/B)		
	Cont	Disc	SARI	Cont	Disc	SARI	Cont	Disc	SARI
LB	0.41	0.41	0.50	0.44	0.26	0.59	0.44	0.33	0.55
SB	0.36	0.52	0.42	0.43	0.22	0.61	0.37	0.45	0.46
KL	0.12	0.84	0.14	0.09	0.66	0.22	0.09	0.79	0.15
SK	0.06	0.46	0.30	0.12	0.84	0.14	0.11	0.93	0.09

[a] Global SARI scores are reported for calculations using potency values for target A or B and selectivity values for target A over B (A/B). Data set identifiers and targets correspond to Table 1. "Cont" and "Disc" stand for continuity score and discontinuity score, respectively. SARI scoring has been found to produce similar results for different molecular representations including, for example, structural key-type and topological fingerprints.[24]

For compound sets LB and SB, both target SARs were heterogeneous, and the L/B and S/B SSRs belonged to the same category. In contrast, for sets KL and SK, all target SARs were globally discontinuous, and the K/L and S/B SSRs were also characterized by strong discontinuity. Thus, the compound data sets for the four cathepsin pairs represented two global SAR categories, and the SAR and SSR phenotypes corresponded for each target pair. Figure 2a shows both potency NSGs and the selectivity NSG for the LB set. The network topology is determined by pairwise compound similarity relationships and is thus common to potency and selectivity NSGs. A characteristic feature of the LB topology is the presence of several distinct sub-graphs or communities of varying potency and selectivity composition indicated by different node colors, which provides evidence of SAR and SSR heterogeneity. In NSG_{LB}, highly selective (red and green) compounds are distributed over different network regions and communities. Figure 2b shows potency and selectivity NSGs for KL. The densely connected networks result from a higher degree of structural homogeneity of the KL than the LB set. The KL topology is characterized by a large central network component including many large nodes, which is characteristic of SAR and SSR discontinuity. However, despite differences in network topology, in NSG_{KL}, highly selective compounds are also found in different local environments, similar to NSG_{LB}.

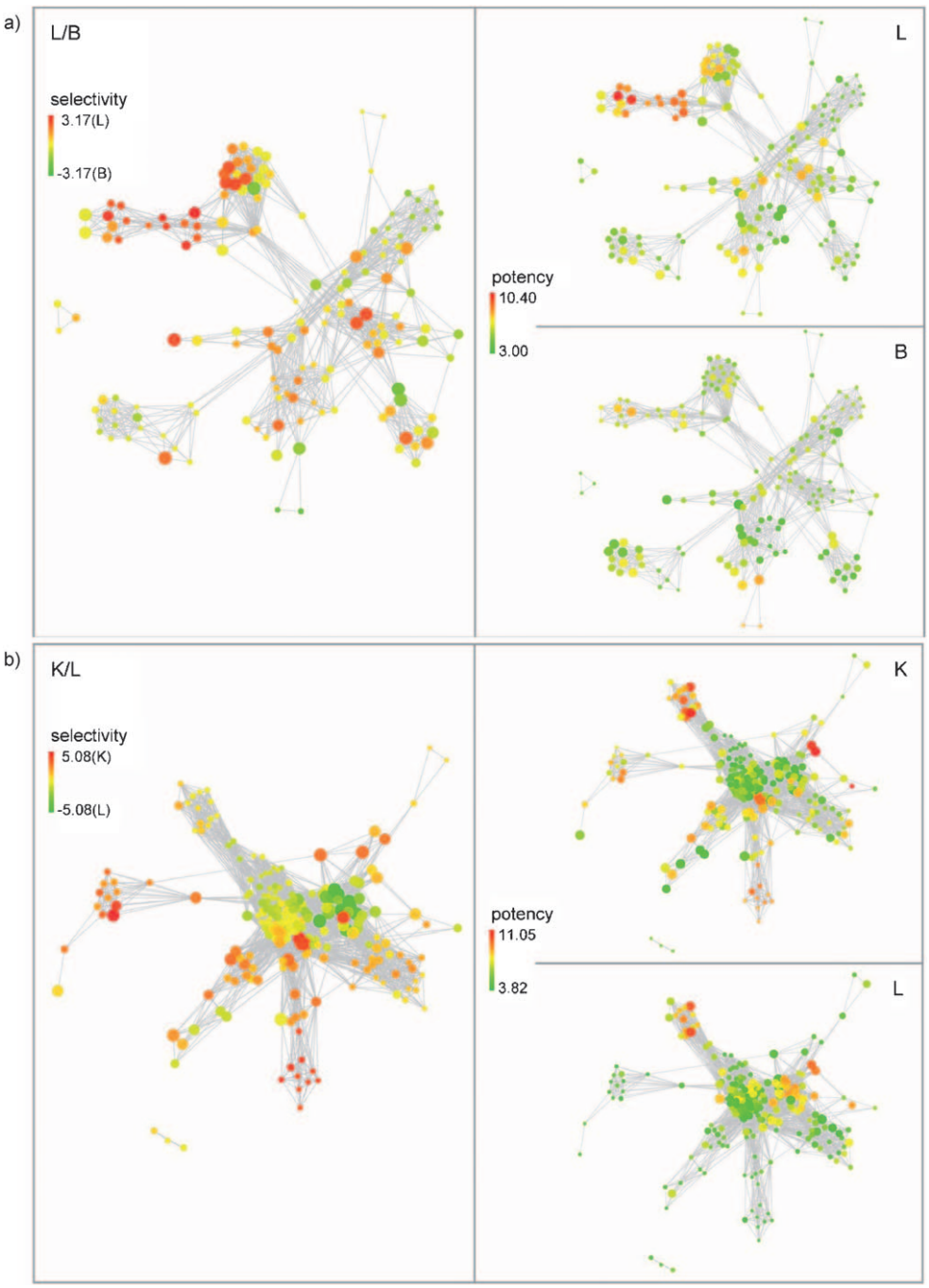

Figure 2. NSG representations for selected target pairs a) L/B, and b) K/L. Each graph at left shows the selectivity NSG_{AB}; at right, the potency NSG_A (top) and NSG_B (bottom) are shown.

Comparison of SAR and SSR features

Comparison of corresponding network segments in potency and selectivity NSGs reveals how compound subsets influence SAR and SSR characteristics. For example, compounds in the upper left clusters in NSG_L in Figure 2a make significant contributions to local SAR discontinuity and global heterogeneity, in contrast to NSG_B, where these compounds are related to each other by a continuous local SAR. Thus, they respond differently to cat L and B. Accordingly, the upper left region of NSG_{LB} reveals that these inhibitors are highly selective and significantly contribute to local SSR discontinuity and global SSR heterogeneity. This is the case because in the same cluster nonselective compounds or compounds selective against the other target are also found. In contrast, compounds in the upper left clusters in NSG_K and NSG_L in Figure 2b include both highly and weakly potent inhibitors that form activity cliffs and make simi-

larly strong contributions to local and global SAR discontinuity. However, in NSG_{KL}, these compounds form a strongly continuous local SSR because they respond to both targets in a very similar way. Thus, these findings illustrate the complementary nature of SAR and SSR information captured in NSG representations.

Cluster discontinuity scores reflect the nature of local SARs and SSRs in compound communities. Figure 3 compares cluster score distributions in potency and selectivity NSGs for the LB and KL sets. The cluster scores essentially cover the entire range of 0 to 1, which reflects a high degree of local SAR and SSR variability. In both cases, SSR heterogeneity is largely determined by one of two targets, L for LB, and K for KL, which produce significantly higher cluster discontinuity scores than their counterparts. The cluster score distributions are overall similar, although the LB SARs and SSR are heterogeneous in nature, whereas the KL SARs and SSR are more discontinuous due to the low degree of chemical diversity, as indicated by the SARI scores in Table 2.

Local SSR environments

After analyzing relationships between global and local SSR features, we now focus on local SSR environments (i.e., SSRs in subsets of similar compounds), ultimately leading to the identification of key compounds and selectivity determinants. Figure 4 shows the selectivity NSGs of the four compound data sets. The comparison further illustrates that NSGs of globally heterogeneous SSRs (Figure 4a and b) are characterized by separate compound communities that reflect a higher degree of chemical diversity in the data sets and distinguish them from NSGs of discontinuous SSRs (Figure 4c and d), which are characterized by a large and densely connected central network component. However, common to all selectivity NSGs is the presence of distinct local SSR environments. Clusters with either very low or high discontinuity scores can be found in each case, regardless of global SSR character. For example, the lower left cluster in Figure 4a (cluster discontinuity score: 0.10) and the cluster at the bottom in Figure 4b (score: 0.01) mostly consist of nonselective inhibitors that make essentially no contributions to SSR discontinuity, represented as small yellow, pale green, or orange nodes. Such environments of local SSR continuity frequently occur in selectivity NSGs and identify compound subsets that provide only little information for the exploration of selectivity at the molecular level. Continuous regions can either be formed by nonselective compounds or by compounds having very similar target selectivity. For example, the cluster at the bottom in Figure 4c (score: 0) and the cluster at the top in Figure 4d (score: 0.14) contain only K-selective compounds (nodes colored in bright red or green, respectively) that are related to each other by continuous local SSRs. Due to their homogeneous selectivity composition, these clusters represent continuous SSR regions, as reflected by the small size of the corresponding nodes.

In contrast, other SSR environments are strongly discontinuous in nature. For example, the cluster on the left in Figure 4b (score: 0.87) contains inhibitors with high (red) and low selectivity (orange/yellow nodes) that make large contributions to

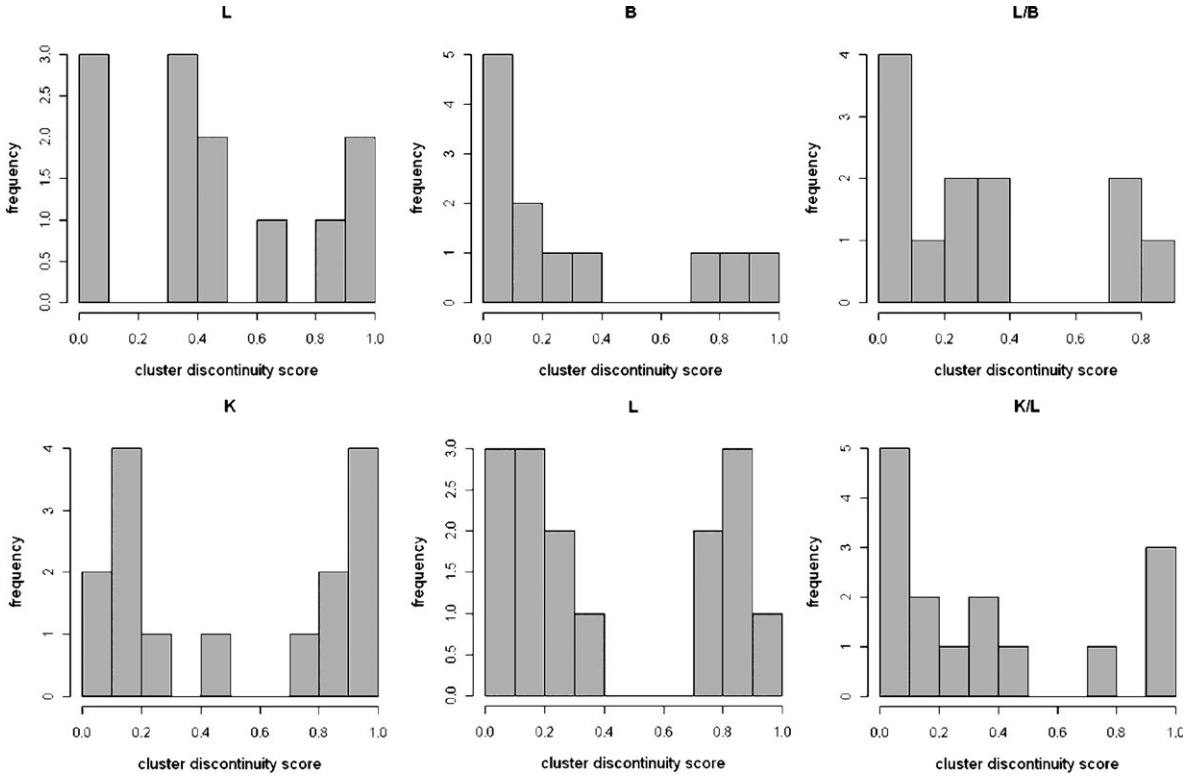

Figure 3. Distribution of cluster discontinuity scores: Histograms of cluster discontinuity score distributions are reported for the LB and KL target pairs. For each pair, three histograms are shown representing the cluster scores in NSG_A, NSG_B, and NSG_{AB}.

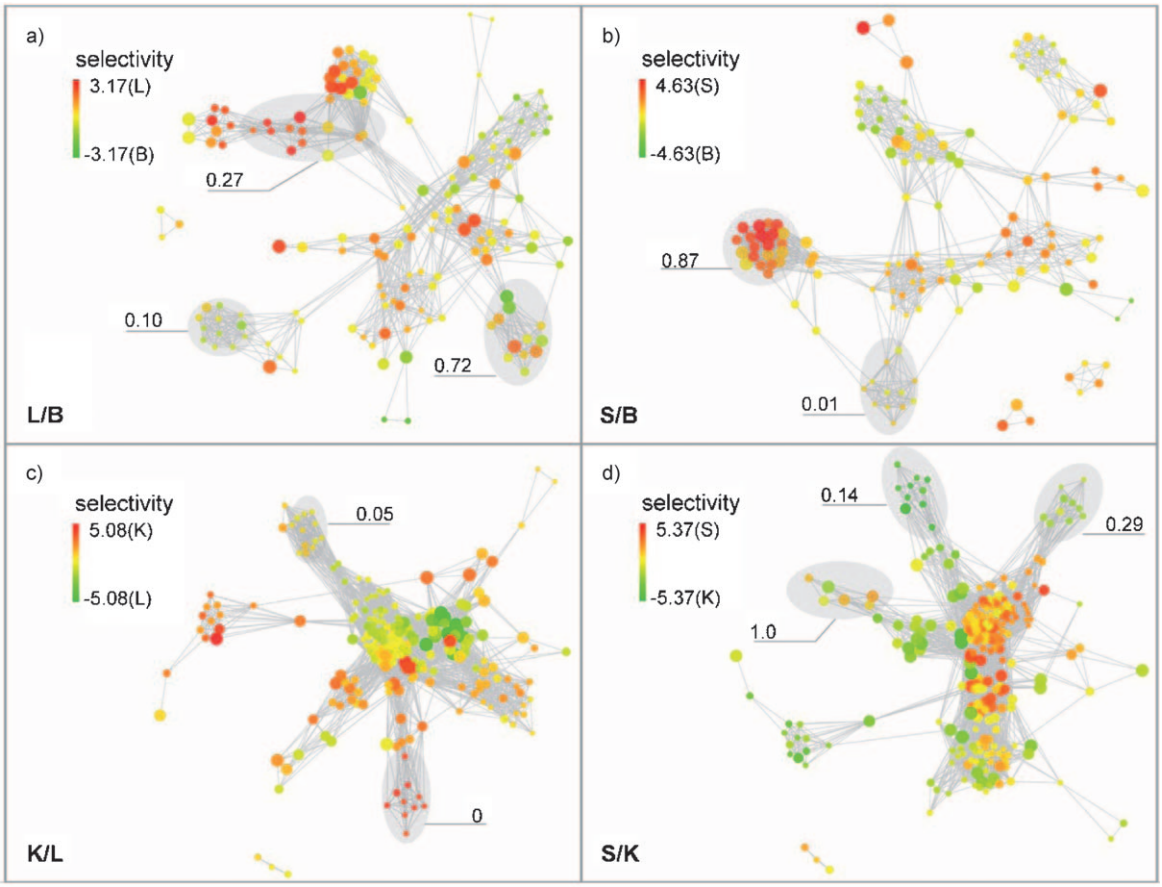

Figure 4. Selectivity NSGs: For each target pair, the selectivity NSG_{AB} is shown. Selected compound clusters are displayed on a gray background and annotated with their cluster discontinuity scores: a) L/B, b) S/B, c) K/L, and d) S/K.

SSR discontinuity. Although many compounds within this cluster have similar selectivity levels, structurally similar selective and nonselective compounds introduce SSR discontinuity and are thus represented by large nodes. Similarly, compounds forming the middle left cluster in Figure 4d (score: 1.0) also strongly contribute to SSR discontinuity. This cluster contains structurally similar inhibitors having significant differences in selectivity, that is, compounds that are either selective for cat S (orange) or cat K (green nodes). Hence, cluster discontinuity scores reveal environments in NSGs that make the largest contributions to SSR discontinuity. The corresponding compound subsets comprise either selective and nonselective inhibitors, or compounds having opposite (i.e., A/B, B/A) selectivity. These clusters generally represent the most interesting NSG regions for the selection of compounds to explore selectivity determinants. Within these environments, compound discontinuity scores provide a measure for the identification of molecules that make key contributions to data-set-specific SSRs.

Activity cliffs, selectivity cliffs, and key compounds

Activity cliffs in potency NSGs are formed by structurally similar compounds with high potency differences, and selectivity cliffs in selectivity NSGs by similar compounds having different selectivity. Hence, most prominent selectivity cliffs are formed by pairs of structural analogues in which one compound is selective for target A and the other is selective for target B (i.e., a pair of large red and green nodes in selectivity NSGs).

Selectivity cliffs are apparent in discontinuous local environments of all selectivity NSGs shown in Figure 4. Compounds with the highest individual SARI discontinuity scores are activity and/or selectivity cliff markers and hence major determinants of SAR and/or SSR features. We selected compounds from NSGs that are activity or selectivity cliff markers and analyzed relationships between cliff markers in potency and selectivity NSGs. Figure 5 shows compounds that are selectivity cliff markers but contribute to single-target SARs in various ways. In Figure 5a, an inhibitor from the LB data set is shown that is selective for cat B and strongly contributes to the formation of selectivity cliffs in NSG_{LB}, having a maximal discontinuity score of 1.0. Moreover, this compound also strongly contributes to local SAR discontinuity in NSG_L and NSG_B with discontinuity scores of 0.96 and 0.81, respectively. In the selectivity NSG_{LB}, this inhibitor is the only cat B-selective compound (green node) within a region containing structurally similar nonselective (yellow nodes) or cat L-selective inhibitors (red nodes). The selectivity for B results from low potency for L ($pK_i = 5.0$) and intermediate potency for B ($pK_i = 6.77$). Thus, this inhibitor has the lowest potency for L and the highest potency for B relative to its neighbors, which explains its contribution to local SAR

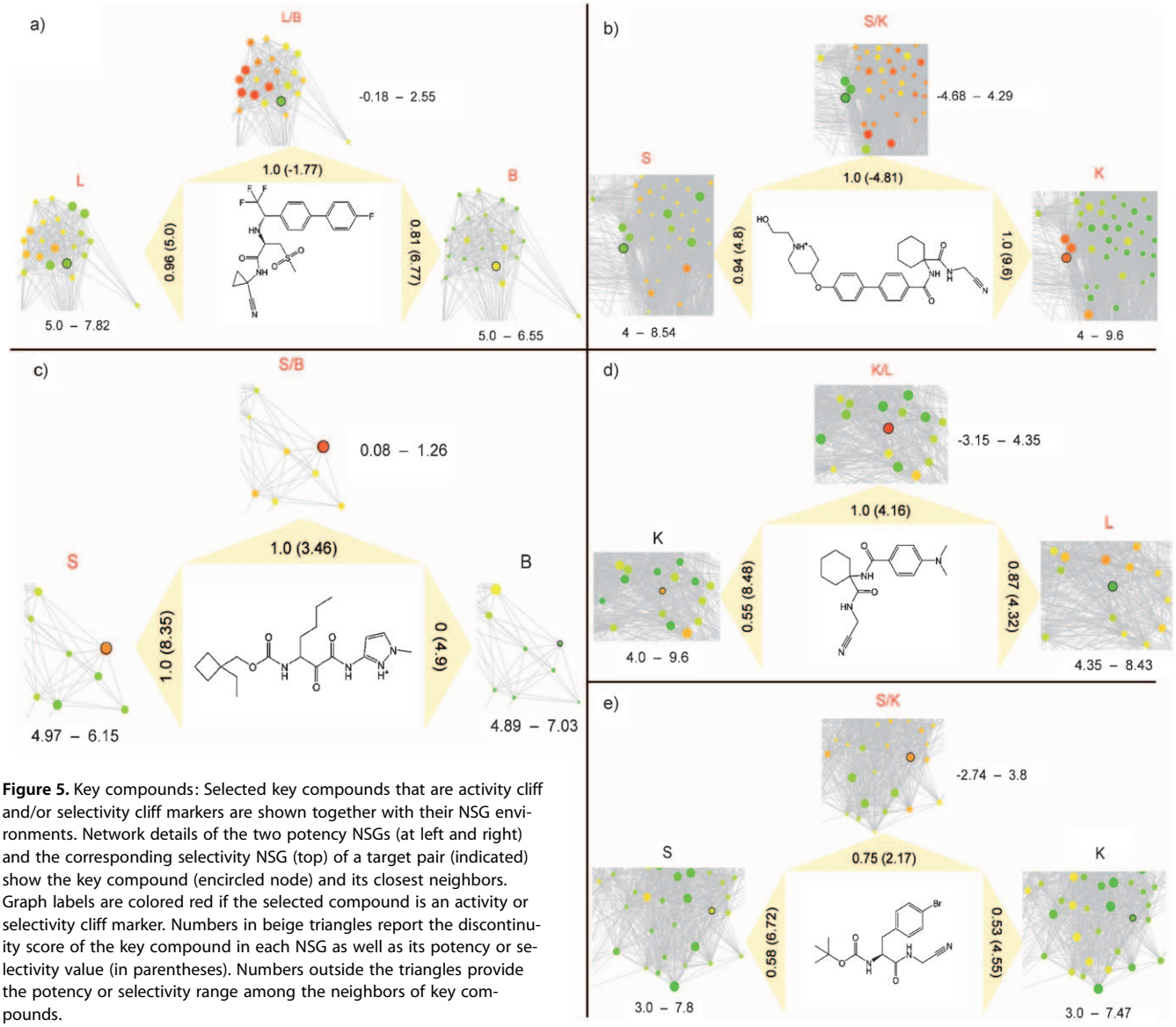

Figure 5. Key compounds: Selected key compounds that are activity cliff and/or selectivity cliff markers are shown together with their NSG environments. Network details of the two potency NSGs (at left and right) and the corresponding selectivity NSG (top) of a target pair (indicated) show the key compound (encircled node) and its closest neighbors. Graph labels are colored red if the selected compound is an activity or selectivity cliff marker. Numbers in beige triangles report the discontinuity score of the key compound in each NSG as well as its potency or selectivity value (in parentheses). Numbers outside the triangles provide the potency or selectivity range among the neighbors of key compounds.

discontinuity in both potency NSGs. Figure 5b shows a prominent activity and selectivity cliff marker for the SK data set. This compound is selective for cat K and participates in the formation of a pronounced selectivity cliff with other K- and S-selective inhibitors that are neighbors in the network, giving rise to strong SSR discontinuity. Similar to the previous example, this key compound is also an activity cliff marker in both potency NSGs due to its low potency for S and high potency for K. However, selectivity cliff markers are not always activity cliff markers as well. Figure 5c shows a compound of the SB data set that is the only S-selective inhibitor in a local environment of nonselective molecules and thus causes strong SSR discontinuity, consistent with its maximal discontinuity score. This compound is highly potent against S, but only weakly potent against B. It is only an activity cliff marker in NSG_S but not in NSG_B because in both local SAR environments, neighboring compounds are only weakly potent. Furthermore, Figure 5d shows an inhibitor from the KL data set that is a prominent selectivity cliff marker because it is highly selective for K, whereas its neighbors are mostly selective for L. The potency of this compound against K is similar to its neighbors, and hence there is no activity cliff in this region of NSG_K. The selectivity of this compound is largely determined by its low potency against L, as illustrated by the complementarity of the node colors in its NSG_{KL} and NSG_L environments. In NSG_L, this compound is responsible for local SAR discontinuity because it has considerably lower potency than its neighbors. Moreover, compounds that do not play a role for individual SARs might also become key compounds in selectivity NSGs. Figure 5e shows an example from the SK set. This inhibitor is moderately potent against S and only weakly potent against K. Its potency values fall into the middle of the potency ranges within its network environments, and hence the compound contributes only little to SAR discontinuity. However, this inhibitor is selective for S, whereas most of its neighbors are selective for K, which induces local SSR discontinuity in NSG_{SK}.

These examples illustrate crucial aspects of potency and selectivity NSG analysis. Key compounds can be readily identified

that are responsible for SAR and/or SSR discontinuity, including activity and/or selectivity cliff markers. Such key compounds are often found to play different roles for SARs and SSRs of target pairs. Furthermore, depending on the potency distributions in compound communities, SAR and SSR features are often highly variable.

Selectivity determinants

In addition to identifying key compounds that are responsible for SAR and SSR discontinuity, another major goal of selectivity NSG analysis with practical utility for medicinal chemistry is the exploration of structural features that determine compound selectivity. This can be accomplished by screening selectivity NSGs for compounds with high similarity and discontinuity values, represented by sets of large connected nodes.

Figure 6 shows sets of analogues and their network environments that reveal selectivity-determining substitutions. In Figure 6a, prominent selectivity cliff markers for the LB data set are shown. The presence of nitrile groups (or other strong nucleophiles) generally represents a hallmark of nonselective cathepsin inhibition. Thus, target selectivity must be determined by other functional groups. Comparison of the three inhibitors in Figure 6a shows that halogenated phenyl substituents at the sulfonyl group render analogues selective for L (indicated by red coloring), whereas the halogenated biphenyl derivative is selective for B (green). Figure 6b shows another pair of analogues in which the bulkiness of a hydrophobic substituent at the quaternary amine is responsible for a change in selectivity from cat K to cat S. In addition, the compound pair in Figure 6c includes an analogue selective for S and a biphenyl derivative that is nonselective. Furthermore, in the series of analogues shown in Figure 6d, various oxygen-containing N substituents at the piperazine ring are observed that determine whether a compound is selective for L (orange nodes on the right) or nonselective (yellow nodes on the left).

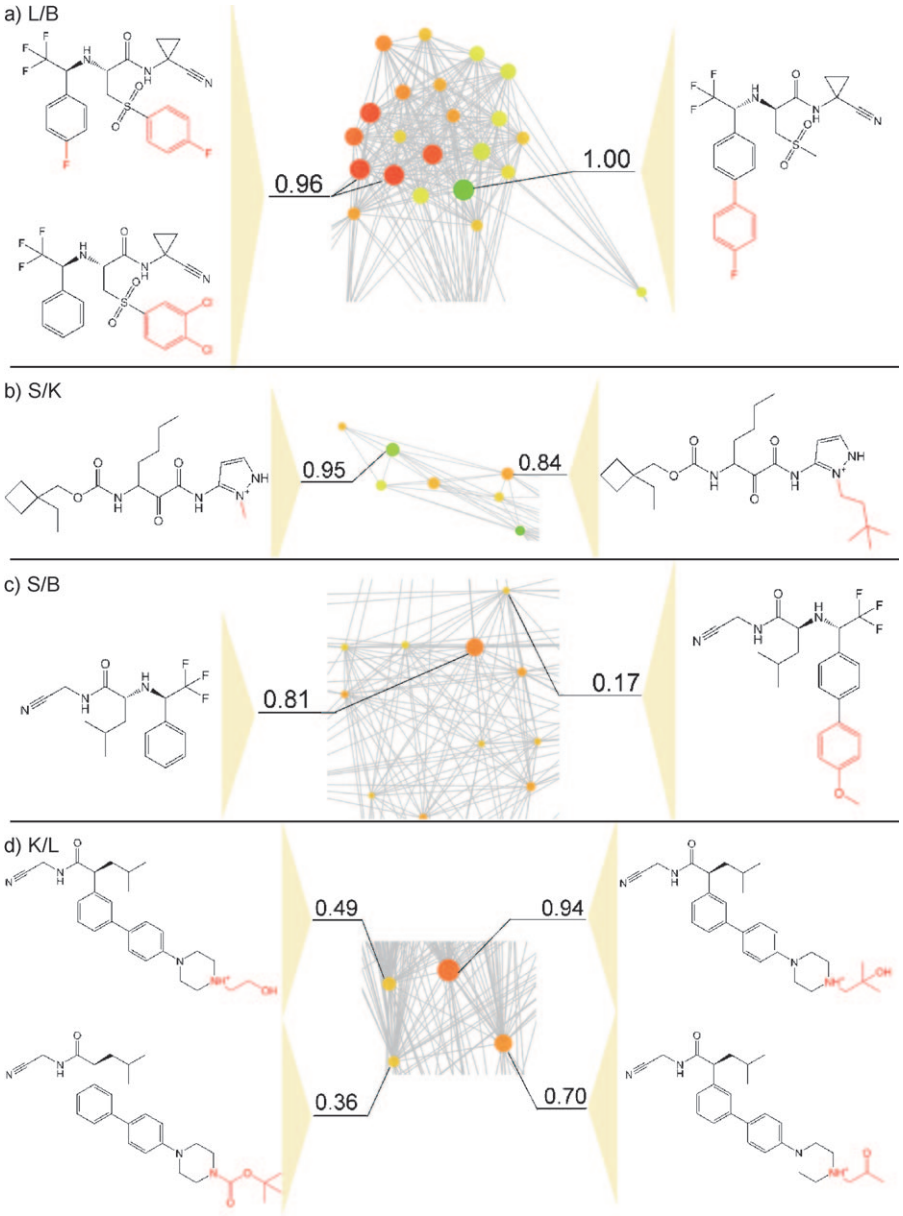

Figure 6. Selectivity determinants: Examples of structurally analogous compounds from all four data sets are shown that form selectivity cliffs of different magnitude. The network environments of these compounds and their discontinuity scores are also displayed. Substituents that distinguish between compounds having different selectivity are colored red.

These results illustrate that series of analogues found in network neighborhoods of key compounds can be used to explore SSRs at the level of individual compounds and to identify selectivity-determining substitution sites and patterns. Thus, selectivity NSG analysis provides global and local SSR information and identifies selectivity determinants.

Conclusions

In this study, we have extended the concept of activity cliffs by introducing selectivity cliffs. Activity cliffs are formed by structurally similar compounds with high potency differences and

can be easily identified in NSG representations. Such activity cliffs can often, but not always, be rationalized on the basis of available X-ray structures of protein–ligand complexes.[23] Going beyond activity cliffs, selectivity cliffs are formed by similar compounds having different selectivity, which results from substantial differences in potency against target pairs, and can be analyzed in selectivity NSGs introduced herein. To quantitatively assess relationships between molecular structure and selectivity, SARI scoring and NSG analysis were combined and applied to sets of inhibitors with varied selectivity against cathepsin targets. Selectivity values were derived from potency differences, and thus structure–selectivity relationships could be globally categorized on the basis of the SARI classification scheme. Heterogeneous and discontinuous SSRs produced different NSG topologies. The analysis of local SSR features consistently identified regions of significant SSR discontinuity containing selectivity cliffs of different magnitude. From these regions, key compounds were selected that influenced SSRs and single-target SARs in similar or different ways, including inhibitors that were activity and/or selectivity cliff markers. From the network neighborhood of selectivity cliffs, series of structurally analogous compounds having high discontinuity scores were isolated, making it possible to identify substitution sites and patterns that were selectivity determinants. For a systematic exploration of structure–selectivity relationships, the ability to identify key compounds and molecular selectivity determinants through selectivity NSG analysis has considerable practical utility.

Acknowledgements

We thank Mathias Wawer for helpful discussions. L.P. is supported by Boehringer Ingelheim, and Y.H. by the B-IT Foundation.

Keywords: activity cliffs · compound potency · molecular networks · structure–activity relationships · target selectivity

[1] G. M. Maggiora, *J. Chem. Inf. Model.* **2006**, *46*, 1535–1535.
[2] L. Peltason, J. Bajorath, *Chem. Biol.* **2007**, *14*, 489–497.
[3] H. Eckert, J. Bajorath, *Drug Discovery Today* **2007**, *12*, 225–233.
[4] H. Kubinyi, *Drug Discovery Today* **1997**, *2*, 457–467.
[5] E. X. Esposito, A. J. Hopfinger, J. D. Madura, *Methods Mol. Biol.* **2004**, *275*, 131–214.
[6] A. L. Hopkins, *Nat. Chem. Biol.* **2008**, *4*, 682–690.
[7] G. V. Paolini, R. H. B. Shapland, W. P. van Hoorn, J. S. Mason, A. L. Hopkins, *Nat. Biotechnol.* **2006**, *24*, 805–815.
[8] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, B. K. Shoichet, *Nat. Biotechnol.* **2007**, *25*, 197–206.
[9] J. Hert, M. J. Keiser, J. J. Irwin, T. Oprea, B. K. Shoichet, *J. Chem. Inf. Model.* **2008**, *48*, 755–765.
[10] M. W. Karaman, S. Herrgard, D. K. Treiber, P. Gallant, C. E. Atteridge, B. T. Campbell, K. W. Chan, P. Ciceri, M. I. Davis, P. T. Edeen, R. Faraoni, M. Floyd, J. P. Hunt, D. J. Lockhart, Z. V. Milanov, M. J. Morrison, G. Pallares, H. K. Patel, S. Pritchard, L. M. Wodicka, P. P. Zarrinkar, *Nature Biotechnol.* **2008**, *26*, 127–132.
[11] J. Bajorath, *Curr. Opin. Chem. Biol.* **2008**, *12*, 352–358.
[12] L. Peltason, J. Bajorath, *J. Med. Chem.* **2007**, *50*, 5571–5578.
[13] R. Guha, J. H. Van Drie, *J. Chem. Inf. Model* **2008**, *48*, 646–658.
[14] J. Bajorath, L. Peltason, M. Wawer, R. Guha, M. S. Lajiness, J. H. Van Drie, *Drug Discovery Today* **2009**, *14*, 698–705.
[15] M. Wawer, L. Peltason, N. Weskamp, A. Teckentrup, J. Bajorath, *J. Med. Chem.* **2008**, *51*, 6075–6084.
[16] D. Stumpfe, H. Geppert, J. Bajorath, *Chem. Biol. Drug Des.* **2008**, *71*, 518–528.
[17] P. Willett, J. M. Barnard, G. M. Downs, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
[18] MACCS Structural Keys: Symyx Software, San Ramon, CA (USA).
[19] T. M. J. Fruchterman, E. M. Reingold, *Software—Practice and Experience* **1991**, *21*, 1129–1164.
[20] R Development Core Team, "R: A Language and Environment for Statistical Computing", *R Foundation for Statistical Computing*, Vienna (Austria) **2008**.
[21] G. Csardi, T. Nepusz, *InterJournal* **2006**, 1695.
[22] J. H. Ward, *J. Am. Stat. Assoc.* **1963**, 58, 236–244.
[23] M. T. Sisay, L. Peltason, J. Bajorath, *J. Chem. Inf. Model.* **2009**, *49*, in press.
[24] L. Peltason, J. Bajorath, *Future Med. Chem.* **2009**, *1*, 451–466.

Received: July 21, 2009
Revised: August 15, 2009
Published online on September 11, 2009

