International Journal of High Throughput Screening

Dovepress
open access to scientific and medical research

Open Access Full Text Article

METHODOLOGY

Computational characterization of SAR microenvironments in high-throughput screening data

This article was published in the following Dove Press journal:
International Journal of High Throughput Screening
31 March 2010
Number of times this article has been viewed

Mathias Wawer*
Su Sun*
Jürgen Bajorath

Department of Life Science Informatics, Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany; *These authors have contributed equally to this work.

Purpose: A computational approach is described to analyze structure–activity relationship (SAR) information contained in compound and screening data sets. The methodology is designed to explore SAR information in a systematic and compound-centric manner in order to aid in the selection of hits from high-throughput screening (HTS) data.

Methods: Chemical neighborhood graphs integrate a graphical representation of the chemical environment of each active compound in a data set with the potency distribution within its neighborhood and information from a quantitative SAR analysis function. Environments are systematically generated and ranked by SAR information content. From these environments, key compounds and compound series can be selected.

Results: The methodology is described in detail. In addition, the application to four screening data sets is reported, revealing different SAR characteristics. A number of different examples of compound environments are presented and discussed that have varying SAR information content.

Conclusion: Chemical neighborhood graphs provide an intuitive graphical access to SAR information contained in hit sets. SAR information is analyzed in a compound-centric manner, with a focus on local SAR environments (microenvironments). It is anticipated that this approach will complement and help to further refine current hit selection strategies and trigger the development of additional graphical analysis methods to search for SAR information in HTS data.

Keywords: screening data sets, hit selection, computational analysis, graphical representation, structure–activity relationship information

Introduction

Given the large size of current screening libraries, high-throughput screening (HTS) campaigns typically produce large numbers of active compounds, often 0.1%–1% of a screening library. Even after secondary screens and confirmatory assays, hundreds of active compounds, or even more, might remain for further study.[1] The chemical exploration of confirmed hits presents a major bottleneck in early-phase drug discovery because it is of course not possible to build a medicinal chemistry program around each chemotype found to be active in a screening campaign. Simply put, there are usually many more hits to choose from than one could possibly explore in hit-to-lead or lead optimization efforts. Consequently, hit selection becomes a rather critical task in the post-screening phase to effectively bridge HTS and medicinal chemistry programs, which is well recognized in pharmaceutical research.[1] Given the large amount of compound activity data that are accumulating in HTS projects,

Correspondence: Jürgen Bajorath
Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113
Bonn, Germany
Tel +49-228-2699-306
Fax +49-228-2699-341
Email bajorath@bit.uni-bonn.de

selection of hits solely on the basis of manual data analysis, guided by chemical intuition and experience, is virtually impossible.

Clearly, the interface between HTS and discovery chemistry is an area where scientists from different disciplines including biologists, screeners, medicinal chemists, and computational scientists need to closely interact. Computational methods are nowadays routinely applied in order to organize and statistically analyze compound activity data as a prerequisite for hit selection.[2] However, currently there are no standard recipes available for selecting the "most interesting" compounds from screening data; moreover, it is often unclear what "most interesting" means in this context. Simply selecting the most active compounds from a screening data set is not sufficient to provide a sound basis for hit-to-lead efforts because facilitating successful hit-to-lead transitions depends, first and foremost, on establishing sustainable and evolvable structure–activity relationships (SARs). Accordingly, both structural and activity criteria must be taken into account. For this reason, screening hits are usually computationally clustered on the basis of structural similarity.[3,4] Then the distribution of hits in different clusters is analyzed and representative compounds are selected from these clusters, taking into account the number of hits a cluster contains. Such procedures ultimately select subsets of hits that cover the structural spectrum of active compounds, but do not take SAR characteristics into account. Therefore, clusters and/or pre-selected hits are often further analyzed by studying their core structures, or maximum common substructures, in order to produce a chemically intuitive organization of active compounds from which initial SAR information might be deduced. Furthermore, molecular scaffolds of active compounds can also be systematically decomposed on the basis of structural rules and organized in tree structures.[5]

For analyzing hit distributions in screening data, visualization tools have become indispensable[6–8] and for interactive analysis, display items such as scatter plots or activity-based heat maps are routinely used. Recently, molecular network representations have also been utilized to mirror compound similarity and activity relationships in compound and screening data sets.[9] However, none of the approaches that are currently applied to structurally organize active compounds or graphically analyze hit distributions is capable of directly extracting SAR information from screening data sets. Nevertheless, this is a critical task because, as mentioned above, evidence of SAR behavior is a key criterion to estimate the likelihood that selected hits evolve into potent leads (in addition to other criteria; eg, synthetic accessibility, predicted metabolic stability, or potential toxicity).

In order to aid in the extraction of SAR information from screening data, we have attempted to organize structural and activity information contained in data sets in different ways and recently introduced the concept of "SAR pathways" as a computational approach to systematically search for SAR information in hit sets.[10] SAR pathways organize active compounds as sequences of pairwise structurally similar molecules that follow an ascending potency gradient leading to, for example, the most active compounds in a data set. Through systematic mining of SAR pathways, SARs (if present in a data set) can be isolated and visualized.[10] As such, SAR pathways represent one possible concept to extract SAR information from screening sets, although they too have their limitations. For example, because active compounds are organized into molecular sequences on the basis of whole-molecule similarity, different chemotypes can participate in the formation of such pathways, which might make the interpretation of SAR characteristics difficult on occasions.

Here we describe another computational methodology to analyze SAR information contained in hit sets that conceptually differs from SAR pathways because it is compound-centric in nature. Our so-called "chemical neighborhood graphs" (CNGs) analyze the structural neighborhood of each individual active compound in a data set and annotate it with SAR-relevant information. Therefore, this data structure is well-suited to characterize "SAR microenvironments" that are formed by a series of similar active molecules with different potency distributions. CNGs are automatically generated and ranked on the basis of SAR information content. Much emphasis is put on intuitive graphical representations of chemical neighborhoods, interpretability of SAR features, and visualization of key compounds. The basic CNG data structure and display tools are made publicly available in order to support both HTS data analysis and computational method development in the scientific community. Herein the CNG approach is described in detail and applied to analyze four public domain screening data sets to illustrate its key features.

Material and methods

Data sets

Hit sets (enzyme inhibitors) from four screening data sets available in PubChem,[14] as summarized in Table 1, were analyzed in this study. From these sets, only those compounds

Table 1 Summary of PubChem hit sets explored in this study

Target	PubChem AID (compound identifier)	Number of active compounds	Lowest potency [μM]	Highest potency [nM]
Cytochrome P450 3A4 (CYP-3A4)	885	3334	39.8	15.8
Cytochrome P450 2C19 (CYP-2C19)	899	1769	39.8	2.5
17-hydroxysteroid dehydrogenase type 10 (HSD17-10)	893	5619	39.8	125.9
Factor XIIa (FXIIa)	852	146	45.3	10.4

Abbreviations: SAR, structure–activity relationship; CNG, Chemical Neighborhood Graph; HTS, high-throughput screening; IC_{50}, half maximal inhibitory concentration; CYP-3A4, cytochrome P450 isoform 3A4; CYP-2C19, cytochrome P450 isoform 2C19; HSD17-10, 17-hydroxysteroid dehydrogenase type 10; FXIIa, Factor XIIa.

were selected that were annotated as 'active' and had confirmed potency values (IC_{50}) associated with them.

Generation of chemical neighborhood graphs

CNGs are generated as a radial view of the chemical neighborhood of an active compound. Given a specific reference compound, a CNG presents molecules that are structurally similar to this compound in an organized manner based on similarity and potency relationships to the reference molecule. A schematic radial view is shown in Figure 1. Similarity between compounds is calculated as the Tanimoto coefficient[11] (Tc) based on ECFP4 fingerprints implemented in Pipeline Pilot®.[12] These extended connectivity fingerprints monitor layered atom environments in test compounds and serve as molecular descriptors for our analysis.

To delineate the similarity radius of a neighborhood and assign compounds to it, we define a threshold value of 0.4

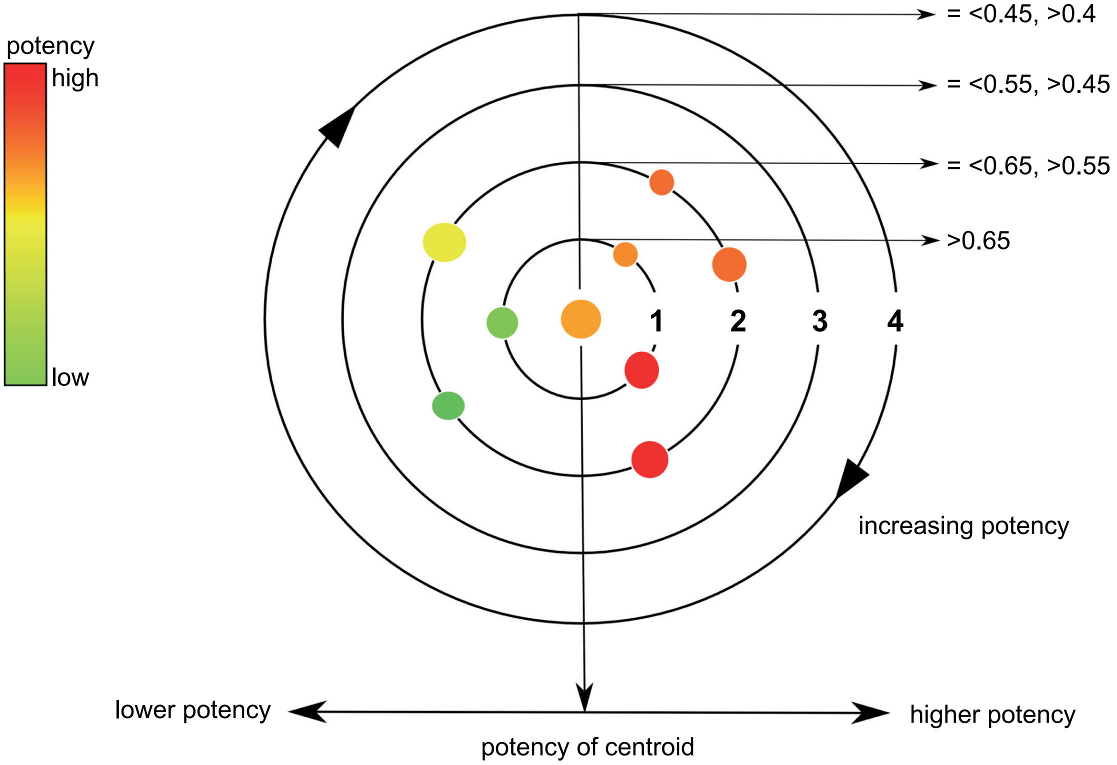


Figure 1 Shown is a schematic drawing of a CNG that illustrates its design. Molecules are represented by colored nodes and the orange node in the center corresponds to the reference compound (central compound) whose neighborhood is shown in the graph. The color code reflects the potency in the data set as indicated in the upper left corner. All compounds whose similarity to the central compound exceeds the predefined (ECFP4 Tc) threshold value of 0.4 are included in the graph representation. They are organized in layers that are defined by non-overlapping similarity ranges and represented by black concentric circles around the center. The similarity decreases from the center to the periphery, as indicated in by the similarity value intervals in the upper right corner. The innermost layer is referred to as the first layer and the outermost as the fourth layer. Compounds having lower potency than the central compound are located on the left half of the CNG and those with higher potency on the right half. In addition, compounds on each layer are arranged in a clockwise manner by increasing potency, as indicated by arrowheads.

for ECFP4-Tc relative to the reference compound as the minimum required similarity; only compounds exceeding this similarity threshold value are included in the neighborhood. Hence, CNGs often capture overlapping subsets of active compounds, which helps to assess and differentiate SAR information content.

The molecules of the neighborhood are arranged in layers of decreasing similarity around the central reference compound and represented as colored nodes. The color code corresponds to compound potency using a continuous color gradient from green to red, corresponding to the lowest and highest potency values present in a data set, respectively.

Nodes that represent compounds with lower potency than the central compound are positioned on the left and more potent compounds on the right half of the graph. On each layer, nodes are arranged in a clockwise manner by increasing potency. Accordingly, the arrangement of nodes on different similarity layers and their color reflects the similarity and potency distribution within the neighborhood of the central compound.

In addition, nodes are scaled in size according to their compound discontinuity score[9] that is calculated for a compound *i* as follows.

$$discontinuity\ (i) = \frac{\sum\limits_{\{j|\mathrm{sim}(i,j)>0.4, i\neq j\}} |\, A_i - A_j \,| \times \mathrm{sim}(i,j)}{\left|\{\, j \mid \mathrm{sim}(i,j) > 0.4,\ \ i \neq j \,\}\right|}$$

Here, A_i and A_j give the logarithmic potency value for compounds *i* and *j*, respectively, and sim(*i*, *j*) corresponds to their ECFP4-Tc similarity value. This discontinuity score is calculated for an active compound by comparing it to all compounds in its environment and scores are normalized with respect to the distribution of scores for the entire data set. Under the assumption of a normal distribution, initial scores are transformed into z-scores for which cumulative probability values are subsequently calculated. The final score thus ranges from 0 (lowest discontinuity) to 1 (highest discontinuity). This discontinuity has a well-defined chemical meaning because it emphasizes structurally similar compounds having large differences in potency. If a compound has many similar neighbors with significant potency differences, it will obtain a high discontinuity score.

Ranking of compound environments

After calculating the CNG for every compound, the resulting environments are independently ranked on the basis of three different parameters in decreasing order:

1. number of compounds within in the environment
2. mean potency of all compounds in the environment
3. discontinuity score of the central compound.

Accordingly, three individual CNG rankings are obtained (one with respect to each parameter) and thus, three rank numbers are assigned to each CNG. The sum of these three rank numbers is calculated and used to generate a final ranking. CNGs are assigned high final ranks if their sum of ranks is low, ie if they achieve high ranks for each of the three parameters. Accordingly, environments that combine a high number of compounds, a high mean potency value, and a high discontinuity score of the central compound are highly ranked.

Workflow

1. Calculate pairwise ECFP4-Tc similarity values for all possible compound pairs in the data set
2. Use each compound in the data set once as the central molecule and identify those compounds whose similarity to the central molecule exceeds the ECFP4-Tc threshold value of 0.4
3. Sort the identified environments by decreasing number of compounds and assign rank numbers
4. Sort the environments by decreasing mean potency and assign rank numbers
5. Sort the environments by decreasing central compound discontinuity score and assign rank numbers
6. For each environment, calculate the sum of these three rank numbers
7. Sort the environments by increasing sum of ranks
8. Display the environments as radial views in the previously determined order

Implementation

Tools to calculate and display CNGs are provided as a Java implementation within the freely available SARANEA program[13] that can be obtained via the following URL: http://www.lifescienceinformatics.uni-bonn.de/0-Seiten/downloads/down.html

For the analysis of CNGs, there are additional convenient graphical tools available in SARANEA that are not described herein. For example, by moving the cursor over a node, the structure of the corresponding compound is displayed, which makes compound selection straightforward.

The calculation of CNGs is not significantly affected by increasing data set size. In addition to increasing numbers of CNGs for larger data sets, the only limiting factor is the calculation of the pairwise compound similarity metric that has quadratic computational complexity. However, this matrix only has to be calculated once.

Results and discussion

CNG data structure and information content

The radial view in Figure 1 summarizes the information contained in CNGs and their organization. Active compounds (nodes) are arranged around the central molecule in four layers of decreasing similarity, each of which corresponds to a defined ECFP4-Tc interval. Nodes are colored according to their potency and organized in the CNG according to their potency relationship with the central compound, ie molecules that are more potent than the central compound are positioned on the right half of the graph and molecules that are less potent on the left half. Furthermore, on each similarity layer, compound potency increases in clockwise direction.

In addition to color-coding, nodes are scaled in size according to the discontinuity score of the corresponding compound, ie the larger the node, the higher the discontinuity score. The discontinuity score conveys important information concerning the contribution of each molecule including the central compound to the SAR discontinuity present within the data set. What does SAR discontinuity specifically refer to? And what would SAR continuity then mean in this context? If a compound is structurally very similar to others within the neighborhood, but has either considerably lower or higher potency, the underlying SAR is discontinuous in nature because there are abrupt changes in activity in the presence of high structural similarity.[9] Such SAR discontinuity is thus characterized by the presence of large neighboring nodes. A particularly interesting case of SAR discontinuity is indicated by the presence of large red and green neighboring nodes, ie structurally very similar compounds having highest and lowest potency within the data set, which form a so-called 'activity cliff'.[9,15] Multiple activity cliffs of different magnitude can be present within a neighborhood. By contrast, combinations of small adjacent and/or distant nodes indicate structurally similar compounds and/or increasingly diverse compounds that on average have only little or moderate differences in activity, which corresponds to SAR continuity.[9] In this case, changes in chemical structure are accompanied by only gradual changes in activity. Moreover, combinations of continuous and discontinuous SAR components can be found within the same data set or microenvironment, thus indicating the presence of SAR heterogeneity.[9,10]

Because an individual CNG is created for each active compound, the number of CNGs available for comparison is typically large, eg a hit set containing 500 compounds yields 500 CNGs; already too many for side-by-side comparisons. Therefore, ranking of CNGs according to SAR information content is an important part of our analysis. In CNG ranking, general criteria are applied, and equally weighted, that are a prerequisite for rich SAR information. The criteria we utilize herein are simple and intuitive and include the number of active compounds within an environment (because small numbers of compounds cannot convey much information), the mean potency of these compounds (because we rather focus on highly than weakly potent compounds), and the discontinuity score of the central compound. Applying this third criterion is particularly important for our analysis because of its compound-centric view and comprehensive nature. Neighborhoods containing SAR discontinuity are *a priori* more rich in SAR information than purely continuous environments. Hence, if we know that a central compound induces SAR discontinuity, the neighborhood is of interest to us. However, we do not wish to primarily focus on neighborhoods that are largely discontinuous in their SAR character so that we do not lose information from

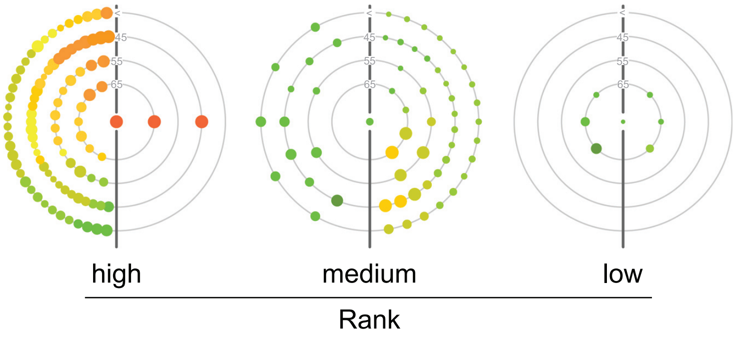
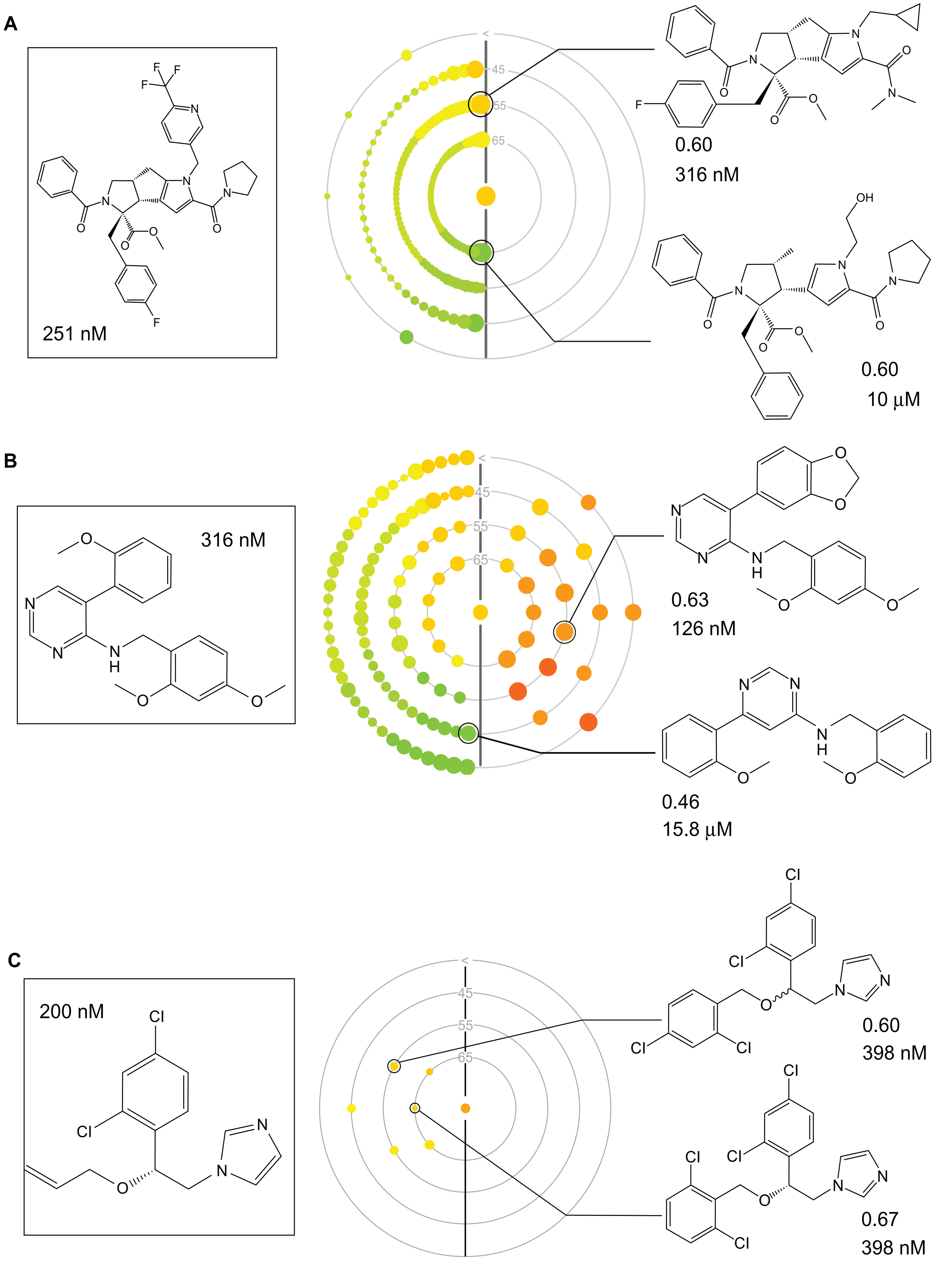


Figure 2 Exemplary CNGs taken from the CYP-3A4 data set (see Table 1) are shown to illustrate the results of the ranking procedure.


A
251 nM
45
55
65
0.60
316 nM
0.60
10 μM
B
316 nM
45
55
65
0.63
126 nM
0.46
15.8 μM
C
200 nM
45
55
65
0.60
398 nM
0.67
398 nM

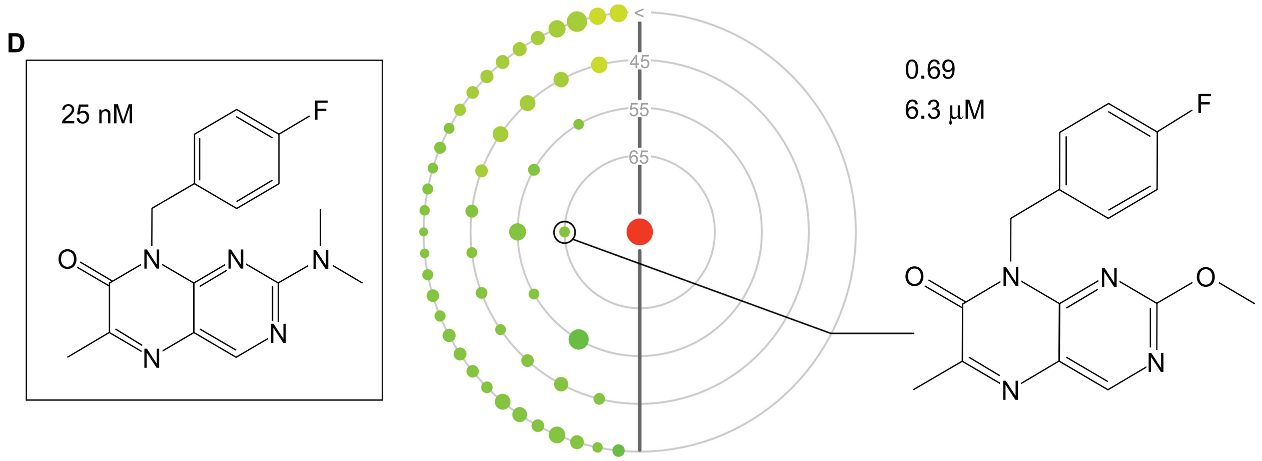


Figure 3 (opposite and above) Four CNGs from the CYP-3A4 data set are shown.The structure and potency of the central compound are always given left to the graph, surrounded by a box. Selected compounds are marked by black circles and their structures, similarity to the central compound, and potency values are shown right to the graph.

heterogeneous environments. Therefore, we do not apply, for example, the mean discontinuity score of a neighborhood as a ranking criterion, but rather the discontinuity score of the central compound.

Taken together, the three applied ranking criteria capture different aspects that are relevant for SAR information and ultimately provide a 'compromise' solution. This scheme generally produces robust and intuitive rankings of CNGs, as illustrated in Figure 2; comparing a highly-ranked (a), mid-range (b) and lowly-ranked (c) compound neighborhood. As can be seen, the number of compounds decreases from (a)–(c), node colors shift towards green (ie lower potency), and the ratio of large to small nodes also decreases, owing to the fact that a high discontinuity score of the central compound means that there usually are at least one or a few large nodes present (that are similar to the central compound but either much more or less active). The ranking of CNGs of a data set makes it not only possible to immediately select top-scoring CNGs but also "scroll" through neighborhoods of gradually changing SAR character and hence recognize patterns that might aid in the selection of additional CNGs for further analysis.

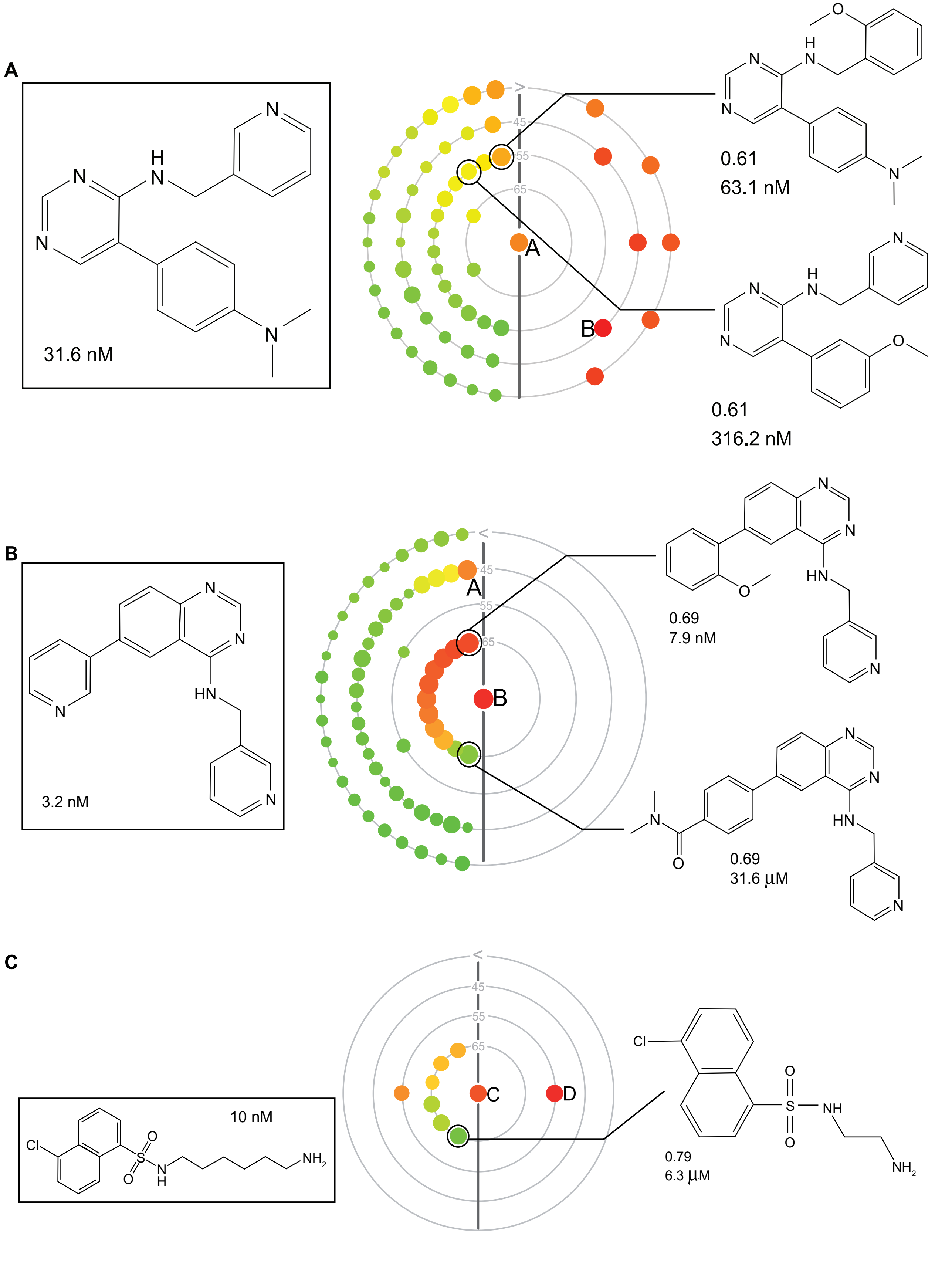
Application to screening data

We have applied the CNG methodology to four hit sets from screens against different types of targets that are publicly available in PubChem. The targets screened were (1) cytochrome P450 isoform 3A4 (CYP-3A4) and (2) isoform 2C19 (CYP-2C19), (3) 17-hydroxysteroid dehydrogenase type 10 (HSD17–10) and (4) factor XIIa (FXIIa). These data sets were selected because they exhibit different types of SAR characteristics.

Cytochrome P450 3A4

Many diverse chemotypes are found in the hit set for target CYP-3A4 and a few of them comprise large numbers of compounds. Accordingly, highly ranked environments are predominantly found to consist of large series of closely related compounds having, however, different potency. The top-ranked neighborhood is shown in Figure 3a. The majority of the 142 compounds contained in this CNG have low to medium potency, as indicated by the dominance of green and yellow nodes. Only a few molecules including the central compound display higher potency, represented as orange nodes. Most compounds are located on the first and second similarity layer indicating that they belong to a series of analogs. With an IC_{50} value of 251 nM, the central compound is the most potent one within this environment and forms moderately-sized activity cliffs with surrounding weakly potent neighbors. The presence of SAR discontinuity in the series including the central compound is due to the presence of only a few key compounds that have significantly higher potency than the majority of molecules in this series. Nevertheless, the dense coverage of the low to medium potency range together with the presence of a few activity cliffs indicate that this compound series is generally amenable to chemical modification leading to an increase in potency. Thus, the central compound and its neighbors might be selected from this data set for further chemical exploration.

A different situation can be observed in the neighborhood shown in Figure 3b, which corresponds to the 13th-ranked CNG and consists of 112 compounds, an example of another well-represented compound series. With an IC_{50} of 316 nM, the central compound has comparable potency to the central compound shown in Figure 3a but has a lower discontinuity


A
31.6 nM
45
55
65
A
B
0.61
63.1 nM
0.61
316.2 nM
B
3.2 nM
45
55
65
A
B
0.69
7.9 nM
0.69
31.6 µM
C
10 nM
45
55
65
C
D
0.79
6.3 µM

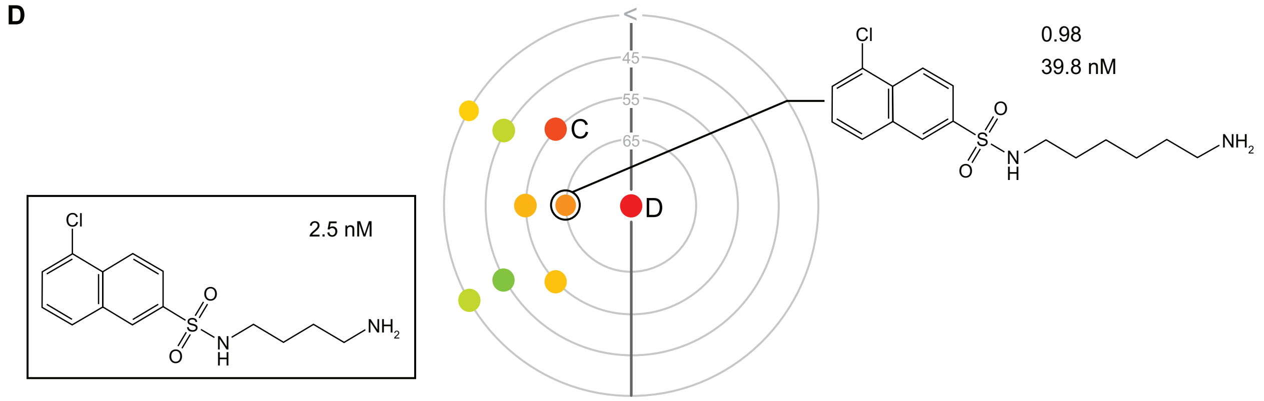


Figure 4 (opposite and above) Four CNGs from the CYP-2C19 data set are shown. The representation in this and the following figures corresponds to the one in Figure 3. The positions of four selected compounds labeled A, B, C, and D are indicated in the graphs.

score. This can be rationalized by examining the potency distribution within the neighborhood. Compared to Figure 3a, more compounds have potency values comparable to the central compound and thus lower its relevance for introducing SAR discontinuity and activity cliffs. The presence of SAR discontinuity is also notable here but it is rather a general feature of this environment and results from contributions of multiple compounds. Accordingly, activity cliffs are formed by different pairs of compounds. This neighborhood obtains a lower score than the top-ranked CNG because it contains significantly fewer compounds and the discontinuity score of the central compound is lower. However, it is much richer in overall SAR discontinuity than the top-scoring neighborhood and also contains continuous SAR elements. Consequently, the series including the central compound in Figure 3b also represents an attractive candidate for selection and further exploration.

Besides the typical appearance of CNGs on medium or low ranks, as illustrated in Figure 2, some lower-ranked neighborhoods are found to display unusual potency distributions. For example, in Figure 3c, a neighborhood consisting of only seven compounds is shown. However, the potency distribution is unusual in that all of these compounds have medium to high potency. Accordingly, there is only little SAR discontinuity and the nodes are small. Moreover, these compounds are structurally not similar to other active molecules. Observing such an environment is a rather rare event. At first glance, this phenotype might be indicative of artificial inhibition or the presence of "frequent hitters". However, the central compound, the fungicide imazalil (PubChem CID 6604394), has been tested in a total of 75 PubChem screening assays, but found to be active in only five of them: screens for substrates or inhibitors of cytochrome P450 isoforms. Moreover, all seven compounds in this environment belong to the imidazole family of fungicides, known inhibitors of cytochrome P450 isoforms,[15] and their mechanism of action depends on interfering with ergosterol synthesis by inhibiting fungal P450 enzymes.[17,18] The unusual potency distribution observed in Figure 3c is therefore unlikely to represent an artifact, but might rather be attributed to screening set composition bias.

Another unusual situation is evident in the CNG shown in Figure 3d. Here the central compound has very high potency (25 nM) but its structural neighbors are all only weakly potent. The central compound is thus the only cause of SAR discontinuity within its environment. This phenotype might be indicative, for example, of a false-positive measurement, but searching PubChem does not provide conclusive evidence. Although this compound was found to show auto-fluorescence, emitted at ~350 nM (PubChem AID 589 and 590), the cytochrome assays studied here used a luciferin read-out, detecting luminescence at about 562 nM.[19] Thus, the measurement leading to the phenotype in Figure 3d might be correct. Nevertheless, SAR information is unusually sparse within its neighborhood, which obtains only a low rank (831) in this data set, and the central compound might be flagged as a potential 'outlier' and one would need to be cautious prioritizing it only on the basis of its apparent high potency.

Cytochrome P450 2C19

Similar to the CYP-3A4 screen, highly ranked CNGs for the CYP-2C19 isoform are dominated by only a few chemotypes, as illustrated in Figure 4a and 4b. However, in contrast to the previously discussed P450 isoform hit set, compounds with high potency are more frequent in these series, although neighborhoods generally contain fewer compounds.

The CNGs shown in Figure 4a and 4b are ranked second and 16th, respectively, and were selected because their central compounds (referred to as A in Figure 4a and B in 4b) are structurally similar and represent two prevalent series. Consequently, their neighborhoods overlap, notably by sharing highly potent compounds. The red nodes seen on the right side of the graph in Figure 4a correspond to the red nodes on the first similarity layer in Figure 4b, thus indicating that they are more similar to series B. The main structural difference between these two central compounds, which alters the composition of their neighborhoods, is the central ring system that consists either of a pyrimidine-4-amine (A) or quinazoline-4-amine (B). Both neighborhoods contain only derivatives of these two ring systems. The red nodes on the first layer in Figure 4b represent close analogs that contain the central quinazoline-4-amine substructure and, in addition, have a pyridine-3-methyl group bound to the amine. The combination of these two substructures is not found in any other molecule of the neighborhood. By contrast, the yellow and orange nodes on the third layer in Figure 4b correspond to compounds containing a central pyrimidine-4-amine group substituted with a pyridine-3-methyl at the amine. All compounds containing a quinazoline-4-amine ring system are either highly or weakly potent, which leads to high SAR discontinuity. However, several compounds containing the central pyrimidine-4-amine structure have medium potency, which indicates that these two series are characterized by different SARs. The quinazoline derivatives have higher potency values than the pyrimidines which are due to the presence of the additional pyridine ring. Some of the more potent pyrimidine compounds also carry this substituent. The presence of additional nitrogen atoms in all medium to highly potent compounds might suggest a pharmacophore resemblance that is not obvious by focusing on the individual compound series. Thus, the comparison of the two ring systems found in these overlapping neighborhoods reveals detailed SAR information and suggests a practical analog design strategy. Figure 4c and 4d show two CNGs representing the same compound series that are differently arranged due to the alternative selection of the central compound.

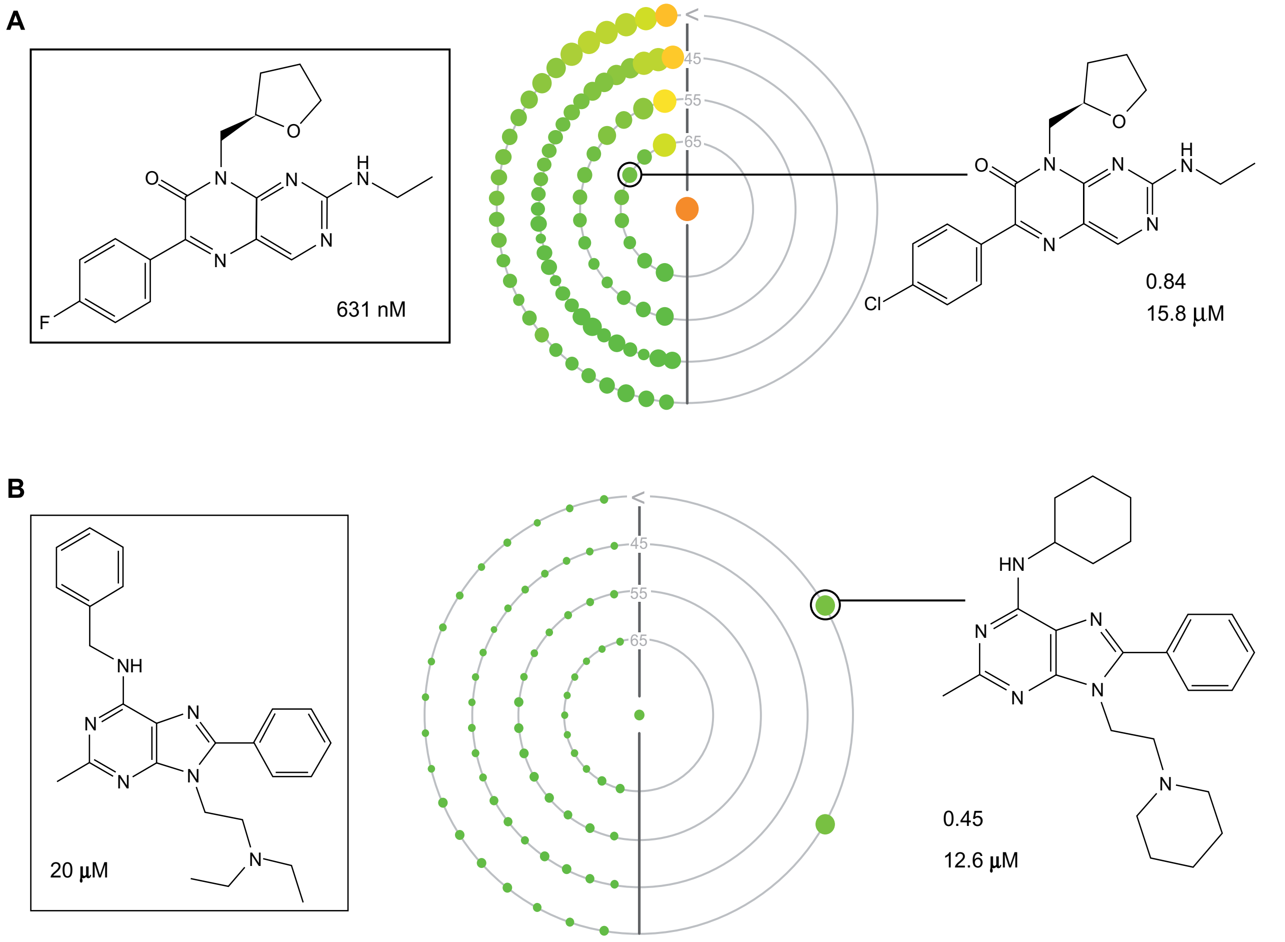


Figure 5 Two CNGs from the HSD17–10 data set.

When choosing compound C as the center (Figure 4c), all compounds are found in the first and second similarity layer, whereas the compounds are distributed over all four layers when compound D becomes the central compound (Figure 4d). Hence, compound C is more similar to other compounds in this series than compound D. Both compounds are highly potent (C: 10 nM, D: 2.5 nM), but compound C is structurally very similar to three weakly potent compounds and thus forms steep activity cliffs, whereas compound D is more similar to compounds having higher potency. Thus, in this case, bioactivity better correlates with structural similarity to compound D, which represents a continuous SAR component. By contrast, compound C introduces SAR discontinuity. These different relationships might also be exploited in analog design.

17-hydroxysteroid dehydrogenase type 10

Compared to the examples discussed thus far, highly potent compounds are rare in the HSD17–10 data set. Although many active compounds were found in this screen, the potency distribution is narrower than observed for the cytochrome P450 isoform hit sets (see also Table 1) and compounds with low potency dominate the HSD17–10 set. Consequently, compound neighborhoods generated from these hits contain many green nodes. Figure 5a shows the top-ranked CNG. SAR discontinuity is mainly introduced by four compounds with medium to high potency and there is only limited SAR information contained in this neighborhood. For this data set, the main source of SAR information is the analysis of activity cliffs. For the top-ranked environments of the HSD17–10 hit set, the distribution of compounds and potency values is comparable to the example shown in Figure 5a. Thus, although this screening set contains only limited SAR information, multiple moderately sized activity cliffs are detected in high scoring neighborhoods that can aid in compound selection. This information would not be available by studying potency value distribution alone.

The neighborhood depicted in Figure 5b contains a compound series that is well-represented among the screening hits. All 79 compounds in this neighborhood are purine derivatives. This CNG represents a prototypic

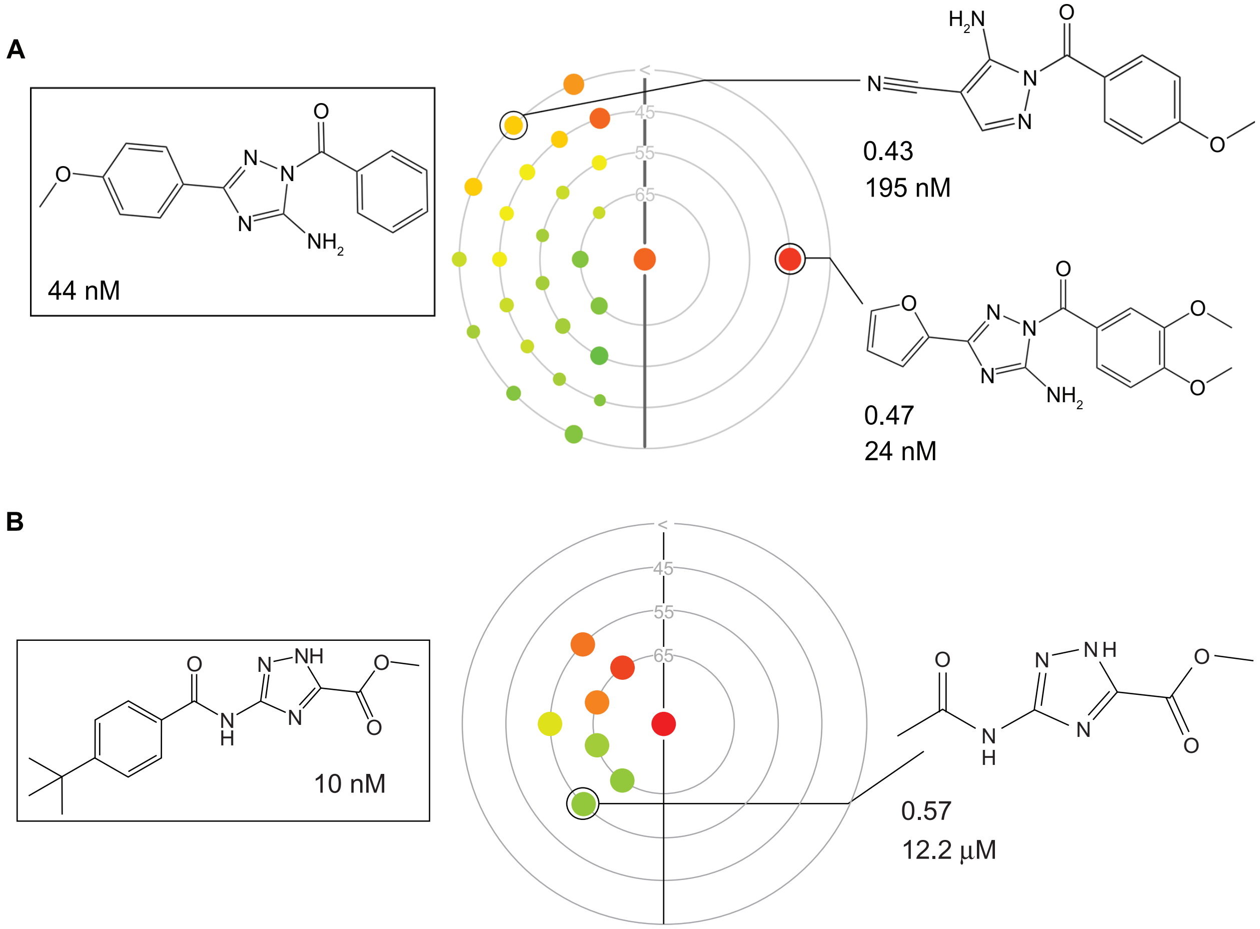


Figure 6 Two CNGs from the FXIIa data set.

example of what would be considered a 'flat' SAR. All compounds have virtually the same low potency (~30 µM) but span the entire similarity range of the neighborhood. Thus, structurally increasingly dissimilar compounds retain similar potency, which indicates that further chemical modifications might not yield more potent compounds (corresponding to 'flat' SAR behavior). Only two exceptions are found on the right side of the CNG that have slightly higher potency values of ~13 µM. Thus, this neighborhood is of very low priority for SAR exploration, as is reflected by its low rank (1975).

Factor XIIa

The set of FXIIa inhibitors deviates from the other examples because it is much smaller in size, consisting of only 146 compounds. However, the majority of these compounds are structurally rather similar. In Figure 6, neighborhoods containing two similar compound series are shown (series A in 6a and B in 6b), both consisting of 1,2,4-triazole derivatives carrying a nitrogen substituent (the only exception is shown in Figure 6a). Nevertheless, these neighborhoods do not overlap. The difference between the series lies in the distribution of other substituents. In series A, the nitrogen substituent is always a primary amine, whereas series B contains an amide linker with varying substituents. In addition, all members of series B have a carboxyl methyl ester group attached to the triazole ring.

Series A in Figure 6a represents the top-ranked neighborhood, contains more compounds than series B (rank 4), and displays a larger potency value spread, although the potency range covered by both series is comparable. Series B in Figure 6b represents a much more discontinuous SAR than series A, which is essentially due the absence of compounds in series B that have medium potency. One would thus preferably compare the SAR characteristics of series A and B after adding more analogues to B. If differences in the potency distribution prevail, distinct SAR characteristics would be associated with these similar triazole series. If not, SAR information might be transferable from one series to the other by comparing their substitution patterns. Hence, comparison of compound neighborhoods also yields differentiated SAR information for similar compound series.

Conclusion

We have introduced chemical neighborhood graphs for the detailed exploration of SAR information contained in screening data. The approach is data-driven and compound centric in nature and methodologically distinct from other computational HTS analysis methods. The CNG method puts much emphasis on studying and comparing SAR microenvironments in a graphical and intuitive manner. Currently, compound potency and discontinuity scores are utilized as compound attributes. However, other parameters such as drug-likeness or synthetic accessibility could be readily added as additional node annotations. Using publicly available screening data, we have demonstrated that different levels of SAR information can be readily extracted from hit sets of different composition and characteristics. This information helps to prioritize compound series for selection and further chemical exploration. The CNG tools are made publicly available to support HTS data analysis and catalyze further methodological developments.

Disclosures

The authors report no conflicts of interest in this work.

References

1. Gribbon P, Lyons R, Laflin P, et al. Evaluating Real-Life High-Throughput Screening Data. *J Biomol Screen*. 2005;10:99–107.
2. Malo N, Hanley JA, Cerquozzi S, Pelletier J, Nadon R. Statistical Practice in High-Throughput Data Analysis. *Nat Biotechnol*. 2006;24:167–175.
3. Stahl M, Mauser H. Database clustering with a combination of fingerprint and maximum common substructure methods. *J Chem Inf Model*. 2005;45:542–548.
4. Böcker A. Toward an improved clustering of large data sets using maximum common substructures and topological fingerprints. *J Chem Inf Model*. 2008;48:2097–2107.
5. Schuffenhauer A, Ertl P, Roggo S, Wetzel S, Koch MA, Waldmann H. The scaffold tree – visualization of the scaffold universe by hierarchical scaffold classification. *J Chem Inf Model*. 2007;47:47–58.
6. Ahlberg C. Visual Exploration of HTS Databases: Bridging the Gap between Chemistry and Biology. *Drug Discov Today*. 1999;4: 370–376.
7. Roberts G, Myatt GJ, Johnson WP, Cross KP, Blower PE Jr. Lead Scope: Software for Exploring Large Sets of Screening Data. *J Chem Inf Comput Sci*. 2000;40:1302–1314.
8. Kibbey C, Calvet A. Molecular Property eXplorer: a novel approach to visualizing SAR using tree-maps and heatmaps. *J Chem Inf Model*. 2005;45:523–532.
9. 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.
10. Wawer M, Bajorath J. Systematic Extraction of Structure–Activity Relationship Information from Biological Screening Data. *ChemMedChem*. 2009;4:1431–1438.
11. Willett P, Barnard JM, Downs GM. Chemical Similarity Searching. *J Chem Inf Comput Sci*. 1998;38:983–996.
12. SciTegic Pipeline Pilot Student Edition, version 6.1.5. San Diego, CA, USA: Accelrys Inc.; 2007.
13. Lounkine E, Wawer M, Wassermann AM, 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(1):68–78.
14. PubChem [website on the Internet]. Bethesda: National Center for Biotechnology Information; c2004 [updated 2009 Nov 23; cited 2009 Dec 9]. Available from: http://pubchem.ncbi.nlm.nih.gov/. Accessed Dec 09, 2009.

15. Maggiora GM. On outliers and activity cliffs – why QSAR often disappoints. *J Chem Inf Model*. 2006;46:1535.
16. Sergent T, Dupont I, Jassogne C, et al. CYP1A1 induction and CYP3A4 inhibition by the fungicide imazalil in the human intestinal Caco-2 cells-comparison with other conazole pesticides. *Toxicol Lett*. 2009;184:159–168.
17. Barasch A, Griffin AV. Miconazole revisited: new evidence of antifungal efficacy from laboratory and clinical trials. *Future Microbiol*. 2008;3:265–269.
18. Vanden Bossche H, Marichal P, Willemsens G, et al. Saperconazole: a selective inhibitor of the cytochrome P-450-dependent ergosterol synthesis in Candida albicans, Aspergillus fumigatus and Trichophyton mentagrophytes. *Mycoses*. 1990;33:335–352.
19. Lottspeich F, Engels JW, eds. *Bioanalytik*. 2nd ed. Munich: Elsevier; 2006.