J Comput Aided Mol Des (2014) 28:919–926
DOI 10.1007/s10822-014-9773-8

Design of an activity landscape view taking compound-based feature probabilities into account

Bijun Zhang · Martin Vogt · Jürgen Bajorath

Received: 12 May 2014 / Accepted: 27 June 2014 / Published online: 8 July 2014

Abstract Activity landscapes (ALs) of compound data sets are rationalized as graphical representations that integrate similarity and potency relationships between active compounds. ALs enable the visualization of structure–activity relationship (SAR) information and are thus computational tools of interest for medicinal chemistry. For AL generation, similarity and potency relationships are typically evaluated in a pairwise manner and major AL features are assessed at the level of compound pairs. In this study, we add a conditional probability formalism to AL design that makes it possible to quantify the probability of individual compounds to contribute to characteristic AL features. Making this information graphically accessible in a molecular network-based AL representation is shown to further increase AL information content and helps to quickly focus on SAR-informative compound subsets. This feature probability-based AL variant extends the current spectrum of AL representations for medicinal chemistry applications.

Keywords Activity landscape design · Molecular networks · SAR visualization · Landscape features · Conditional feature probabilities · Per-compound contributions

B. Zhang · M. Vogt · J. Bajorath (✉)
Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, 53113 Bonn, Germany
e-mail: bajorath@bit.uni-bonn.de

Introduction

Activity landscape (AL) models are generally defined as graphical representations that integrate structural similarity and potency relationships between specifically active compounds [1]. They are primarily used to visualize and explore global and local structure–activity relationship (SAR) features of compound data sets. AL representations of rather different design have been introduced [1, 2] including 2D plots such as structure–activity similarity (SAS) maps [1–3], annotated similarity-based compound networks [1, 2, 4], and 3D models [1, 2, 5]. Regardless of their design, AL representations introduced thus far have in common that they account for pairwise compound similarity and potency relationships [1]. In AL representations, characteristic SAR features can be identified including "activity cliffs" [6–8], "smooth (compound) pairs" [3, 8], and "similarity cliffs" [3, 8]. Activity cliffs are formed by structurally similar (analogous) compounds with large potency differences and are the most prominent features of ALs. By contrast, similarity cliffs are formed by structural distinct compounds having similar potency. In addition, smooth pairs are formed by structurally similar compounds with similar potency. Hence, these three AL features principally represent possible structure/potency relationships between active compounds. Smooth pairs and subsets of similarity cliffs (with moderate structural differences) delineate regions of SAR continuity in ALs in which gradual changes in compound structure are accompanied by gradual changes in potency. By contrast, activity cliffs represent the extreme form of SAR discontinuity in ALs. In discontinuous SAR regions, small structural modifications of active compounds lead to significant potency variations.

Recently, a first approach has been introduced to assign conditional probabilities of landscape features to individual

compounds, rather than compound pairs [8]. Hence, for each compound in a data set, it can be estimated whether it is most likely to participate in the formation of activity cliffs, similarity cliffs, or smooth pairs in a given data set [8]. These probability estimations enable the assignment of individual compounds to SAR feature categories. This statistical framework also provides an opportunity to modulate the compound pair focus of AL design. However, currently no AL representation is available that utilizes this information. Therefore, we introduce herein an AL variant that directly takes SAR feature probabilities for individual compounds into account.

Materials and methods

We first adapt the conditional probability framework for graphical analysis of compound-based AL feature probabilities. Then, we introduce a prototypic AL representation and describe the design of our new AL variant. Finally, compound data used for our analysis are specified.

Conditional probability formalism

The conditional probability formalism was introduced to capture the frequencies of SAR features with respect to a single compound [8]. It is applied to determine the probability of a given compound to form informative SAR features including activity cliffs, similarity cliffs, and smooth pairs. These features can be intuitively rationalized using an SAS map. In an SAS map [3], compound pairs are plotted along two axes, one accounting for structural similarity and the other for potency differences between compounds, as schematically illustrated in Fig. 1. Compound pairs forming different SAR features are located in different quadrants of the map, i.e., activity cliffs in the upper right, smooth pairs in the lower right, and similarity cliffs in the lower left quadrant. To determine the probability of individual compounds to participate in these feature, 'local' SAS maps are generated that only contain all compound pairs formed by a given reference compound [8]. The population density of the individual quadrants then reflects the frequency with which this compound is involved in the formation of different SAR features within a given compound data set. Conditional probabilities are introduced to quantify the propensity of a compound to form activity cliffs in the presence of (conditioned on) another structurally similar compound or the propensity of a compound to form similarity cliffs in the presence of a compound having similar activity. Therefore, two partitions are calculated. Compounds are partitioned according to their potency difference (with respect to the reference compound) into two subsets, ΔPotHigh and ΔPotLow, depending on whether the potency difference to the reference compound does or does not exceed two orders of magnitude (the potency difference criterion for an activity cliff applied herein). In addition, compounds are partitioned according to their structural similarity into two subsets, TcHigh and TcLow, depending on whether the Tanimoto coefficient (Tc) value [9] exceeds a given threshold (similarity criterion). In our calculations, a Tc threshold of 0.55 for the extended connectivity fingerprint with bond diameter 4 (ECFP4) [10] was chosen as the similarity criterion. In a local SAS map, these partitions can be readily visualized, as illustrated in Fig. 1. Here, compounds falling into the blue region correspond to compounds having potency similar to the reference compound (ΔPotLow) while compounds falling into the red region are structurally similar to the reference compound (TcHigh). Using these partitions, smooth pairs, activity cliffs, and similarity cliffs can be characterized. For example, the reference compound forms a smooth pair with each compound that is located in the TcHigh/ΔPotLow quadrant, whereas activity cliffs are formed with the reference by compounds in the TcHigh/ΔPotHigh and similarity cliffs by compounds in the ΔPotLow/TcLow quadrant.

In order to determine the propensity of a reference compound to form activity cliffs, only compounds that are structurally similar to the reference are considered and the fraction of these compounds that also have a large potency difference is determined. Mathematically, this corresponds to the conditional probability:

$$\Pr(\Delta\text{PotHigh}|\text{TcHigh}) = \frac{|\Delta\text{PotHigh} \cap \text{TcHigh}|}{|\text{TcHigh}|} \quad (1)$$

Analogously, the propensity to form similarity cliffs is determined by only considering compounds with potency similar to the reference and determining the fraction of these compounds that also have low structural similarity.

$$\Pr(\text{TcLow}|\Delta\text{PotLow}) = \frac{|\text{TcLow} \cap \Delta\text{PotLow}|}{|\Delta\text{PotLow}|} \quad (2)$$

The complementary probabilities Pr(ΔPotLow|TcHigh) and Pr(TcHigh|ΔPotLow) yield a measure for the probability of a compound to form smooth pairs either with respect to structurally similar compounds or with respect to compounds having similar potencies:

$$\begin{aligned}\Pr(\Delta\text{PotLow}|\text{TcHigh}) &= 1 - \Pr(\Delta\text{PotHigh}|\text{TcHigh}) \\ &= \frac{|\Delta\text{PotLow} \cap \text{TcHigh}|}{|\text{TcHigh}|}\end{aligned} \quad (3)$$

$$\begin{aligned}\Pr(\text{TcHigh}|\Delta\text{PotLow}) &= 1 - \Pr(\text{TcLow}|\Delta\text{PotLow}) \\ &= \frac{|\text{TcHigh} \cap \Delta\text{PotLow}|}{|\Delta\text{PotLow}|}\end{aligned} \quad (4)$$

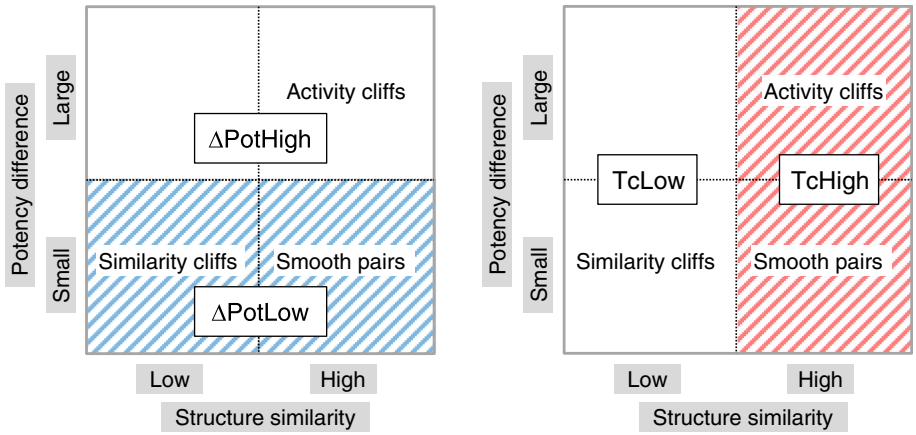


Fig. 1 Local SAS map. In local SAS maps, similarity cliffs, smooth pairs, and activity cliffs are located in the *lower left*, *lower right*, and *upper right quadrants* respectively. Conditional probabilities are determined based (conditioned) on subsets of compounds with small differences in potency (*blue region*, schematic SAS map on the *left*) or subsets of compounds with high structural similarity (*red region*, map on the *right*)

A detailed formal introduction of these conditional probabilities is given in the original work [8]. For probability calculations, fuzzy boundaries between the quadrants defined above were introduced to balance boundary effects [8].

Based on these conditional probabilities, three AL feature categories were selected for the purpose of our current analysis. Compounds were assigned to a given category if the corresponding conditional probability exceeded the pre-defined threshold value (Thresh). We define the following conditional probability-based feature categories (Cat):

Cat 1: similarity cliffs likely:
$\Pr(\text{TcLow}|\Delta\text{PotLow}) \geq \text{Thresh}_1$

Cat 2: smooth pairs likely/activity cliffs unlikely:
$\Pr(\Delta\text{PotLow}|\text{TcHigh}) \geq \text{Thresh}_2$

Cat 3: activity cliffs likely:
$\Pr(\Delta\text{PotHigh}|\text{TcHigh}) \geq \text{Thresh}_3$

The formally possible conditional feature category 'smooth pairs likely/similarity cliffs unlikely' was not considered for AL design. According to our threshold criteria, as explained in detail below, this category would include compounds for which a significant proportion of other compounds (>10 %) having similar potency were also structurally similar. Although of relevance from a statistical point of view, visualization of this category does not further improve the understanding of features in AL representations presented herein. This is the case because relationships primarily based upon similar potency are hard to rationalize in a graph representation based upon structural similarity. Furthermore, for visualization, there is potential ambiguity with respect to category 2 defined above reflecting the propensity to form smooth pairs conditioned on structural similarity. Therefore, the probability of compounds to form smooth pairs was only assigned on the basis of category 2.

The thresholds were set at the 90th percentile of pooled conditional probability values calculated from 148 activity classes taken from the ChEMBL database (release 18) [11]. Each class was required to contain at least 100 compounds spanning a potency range of at least three orders of magnitude. Given the 90th percentile condition, for each category, 10 % of all compounds met or exceeded the respective threshold. From the activity classes, threshold values of $\text{Thresh}_1 = 0.999$, $\text{Thresh}_2 = 1.0$, and $\text{Thresh}_3 = 0.366$ were determined. The high threshold value for category 1 indicated that for ~10 % of the compounds, no or only very few structurally similar compounds with similar potency were available (either because no structurally similar compounds were present in the data sets or, if present, they formed activity cliffs). Likewise, the threshold of 1.0 for category 2 indicated that at least 10 % of all compounds did not form activity cliffs with any structurally similar compound. Feature category assignments were then made for data set compounds on the basis of their calculated probability values for categories 1, 2, and 3. By definition, categories 2 and 3 were mutually exclusive but a compound could, in principle, meet the thresholds for category 1 and 2 or 1 and 3. If a compound qualified for two categories (which was rarely the case), activity cliffs were prioritized over similarity cliffs and similarity cliffs over smooth pairs.

Molecular network-based activity landscape representation

The network-like similarity graph (NSG) [4] is a similarity-based compound network used as an AL representation. In the NSG, nodes are compounds that are connected by edges if their calculated ECFP4 Tanimoto similarity is greater than 0.55. A graphical layout algorithm is applied to organize groups of similar compounds into clusters, separate densely connected clusters from each other, and determine final cluster and node positions [4]. Hence, in the NSG, distances between nodes clusters have no chemical meaning but are determined for layout and visualization

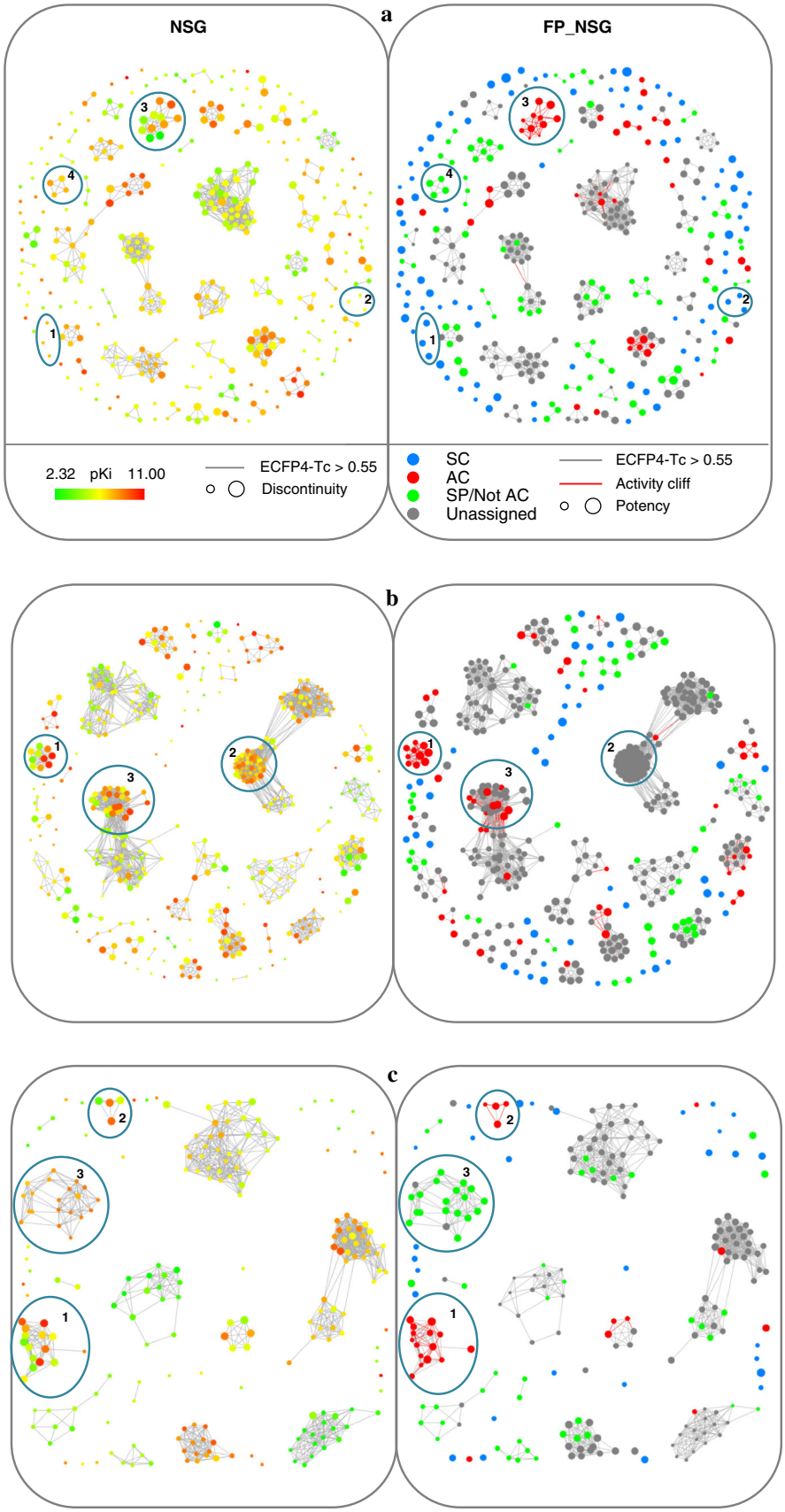
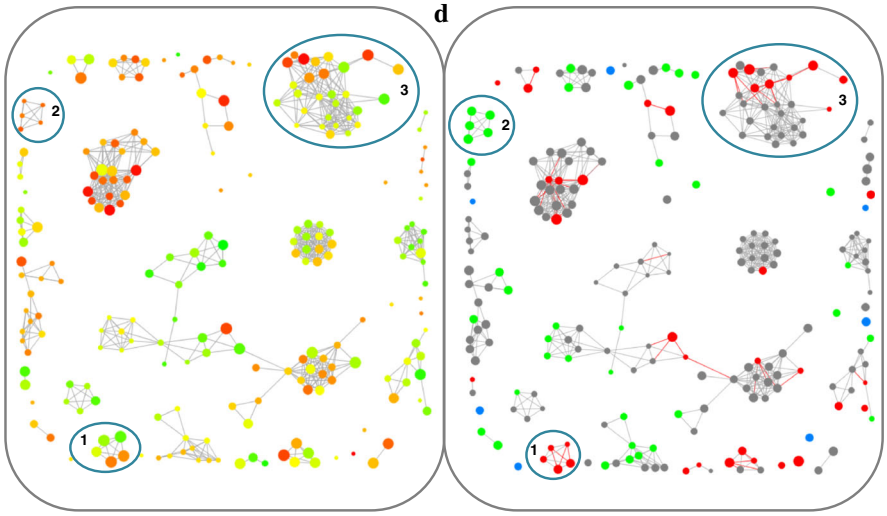


Fig. 2 Activity landscapes. Shown is a side-by-side comparison of NSGs and FP_NSGs for four different compound data sets including ligands of **a** M1, **b**, α-1a, **c** 5-HT3a, and inhibitors of **d** MM3. In the legend of **a**, edge criteria and node annotations are summarized for NSGs and FP_NSGs, respectively. Selected compound subsets and clusters are *encircled* and correspondingly numbered in the NSG and FP_NSG representations

Fig. 2 continued



purposes. Nodes are annotated in two ways. Firstly, they are color-coded according to compound potency using a continuous color spectrum from green (low potency) over yellow to red (high potency). Secondly, nodes are scaled in size according to numerical per-compound compound discontinuity scores [4] that quantify the contribution of a compound to local SAR discontinuity. A compound introduces a high degree of local discontinuity in an AL if its potency significantly differs from the potency values of its structural neighbors (i.e., compounds within the same cluster). The higher the discontinuity score value of a compound is, the larger its node becomes. An exemplary NSG is shown in Fig. 2a (left image).

Activity landscape variant with conditional feature probabilities

Assignment of individual compounds to AL feature categories provided the basis for the design of an NSG variant taking compound feature probability into account. This variant was termed Feature Probability NSG (FP_NSG). In the FP_NSG, the NSG similarity criterion for edge formation between two nodes has been retained. In addition, the network layout has also been conserved. These two criteria enable a direct side-by-side comparison of FP_NSGs and corresponding NSGs. By contrast, in FP_NSGs, compound nodes are colored according to their assigned feature categories:

Cat 1: similarity cliffs: Blue
Cat 2: smooth pairs: Green
Cat 3: activity cliffs: Red

In addition, compounds not assigned to any feature category (unassigned compounds) were colored gray. These compounds do not meet the threshold for any of the three categories and thus have no statistically significant preferential probability to engage in the formation of specific AL features. Such compounds might still carry some SAR information but do not belong to the most characteristic compounds in a given AL.

Activity landscape feature probability-based node coloring represents the most characteristic aspect of FP_NSGs (and replaces the potency-based node coloring in standard NSGs). In addition, in FP_NSGs, nodes were scaled in size by compound potency (replacing the discontinuity score-based node scaling in NSGs). Hence, FP_NSGs are designed to graphically represent information complementary to NSGs. An exemplary FP_NSG is shown in Fig. 2a (right image). All differences in design between NSGs and FP_NSGs, their node annotations, and edge criteria are summarized in Fig. 2a.

Routines to generate and display NSGs and FP_NSGs were implemented in Java with the aid of the OpenEye chemistry toolkit [12].

Data sets

For AL analysis, four exemplary compound activity classes yielding different AL feature distributions were selected from ChEMBL (release 18), as summarized in Table 1. These data sets included ligands of the human muscarinic acetylcholine (M1), alpha-1a adrenergic (α-1a), and serotonin 3a (5-HT3a) receptor and inhibitors of matrix metalloproteinase 3 (MM3). Only compounds with direct interactions against their targets at the highest ChEMBL confidence level were assembled and only equilibrium constants (K_i values) were considered as potency measurements. If multiple K_i measurements were available for a compound, it was only considered if all values fell into the same order of magnitude. In this case, the geometric

Table 1 Data sets and SAR feature statistics

Target	# Compounds	# Similarity cliffs	# Activity cliffs	# Smooth pairs
M1	370	25,575	56	818
α-1a	461	50,955	122	3,046
5-HT3a	225	8,209	33	934
MM3	228	9,866	48	645

For the four data sets used herein, the total number of compounds, similarity cliffs, activity cliffs, and smooth pairs are reported. All possible compound pairs were considered for AL feature analysis

mean of all values was calculated as the final potency annotation.

Results and discussion

Network-like similarity graph representations provide both global views of ALs and local views of specific SAR environments formed by subsets of compounds in data sets. Given potency-based node coloring and SAR discontinuity-based node scaling, activity cliffs appear as combinations of large red and green nodes in NSGs. Furthermore, clusters consisting of small nodes generally represent continuous local SAR environments but cannot be further distinguished. Compound clusters containing prominent activity cliffs (representing strongly discontinuous local SARs) are straightforward to identify in NSGs, which makes it possible to select SAR-informative compound subsets from large and heterogeneous data sets [4].

Figure 2 shows a side-by-side comparison of NSG and FP_NSG representations for different data sets. The comparison reveals complementary information in NSGs and FP_NSGs and an information gain associated with FP_NSGs. In many cases, compound clusters that are essentially indistinguishable on the basis of potency coloring in NSGs are further differentiated in FP_NSGs with respect to AL features. In addition, singletons (i.e., compounds without qualifying similarity relationships to others) are often—but not always—identified as compounds likely to form similarity cliffs in FP_NSGs, as illustrated in Fig. 2a (M1).

Compound subsets forming prominent activity cliffs are easily identified in both NSGs and FP_NSGs as illustrated, for example, by cluster 3 in Fig. 2a or cluster 1 in Fig. 2b (α-1a). Although the NSGs of the M1 and α-1a data sets are similar, Fig. 2b also shows that α-1a contains many more unassigned compounds than M1. Thus, the global SAR information content of these sets differs on the basis of conditional probability assignments. Cluster 2 in Fig. 2b provides a good example of a densely populated region that almost exclusively consists of compounds that are not likely to participate in the formation of well-defined AL features.

The NSG in Fig. 2c (5-HT3a) indicates that this data set is mostly continuous in nature and might contain only little SAR information, with the exception of a prominent cluster (1) and two smaller ones (e.g., 2) that contain activity cliffs. However, the corresponding FP_NSG reveals subsets of compounds with a high probability to form smooth pairs, especially in the prominent cluster 3, or similarity cliffs. By contrast, the NSG in Fig. 2d (MM3) indicates that this data set contains much more information than 5-HT3a. However, the FP_NSG reveals the presence of many unassigned compounds in this data set, in addition to SAR-informative clusters. For example, cluster 3 in Fig. 2d contains prominent activity cliffs but also many structurally similar compounds that are unassigned and hence not likely to form defined AL features.

Taken together, the results in Fig. 2 reveal that the comparison of NSGs and FP_NSGs provides a much more differentiated view of SAR information than is provided by NSGs alone. In FP_NSGs, key compounds representing prominent AL features can be readily identified, which is difficult to accomplish on the basis of NSGs, except for activity cliffs. However, compounds that are unassigned in FP_NSGs might map to interesting local SAR environments and also be considered during SAR analysis.

Figure 3 shows compound examples representing different AL feature categories in FP_NSGs including subsets of compounds likely to form similarity cliffs (Fig. 3a) and compound clusters forming activity cliffs (Fig. 3b) or smooth pairs (Fig. 3c). Given their structures and potency values, these compounds clearly form the AL features that are assigned to them in the FP_NSGs, which reveals a high degree of consistency between per-compound probability assignments and observed AL features. Moreover, Fig. 3b shows an instructive example of differences in information content between NSGs and FP_NSGs and their complementarity. NSGs contain many compound clusters that are characterized by the presence of apparent local SAR discontinuity (Fig. 2) such as, for example, the top cluster in Fig. 3b that can only be further differentiated in the FP_NSG. The cluster comparisons presented above have revealed that not all of these clusters contain activity cliffs. However, the bottom cluster in Fig. 3b provides a prime example for a local SAR environment in which all participating compounds have a high probability to form activity cliffs and should thus be a focal point of SAR analysis. This information can only be obtained in FP_NSGs.

In light of the above, we strongly favor the complementary use of NSGs and FP_NSGs. As demonstrated, FP_NSGs help to further distinguish between clusters representing different local SAR environments and identify

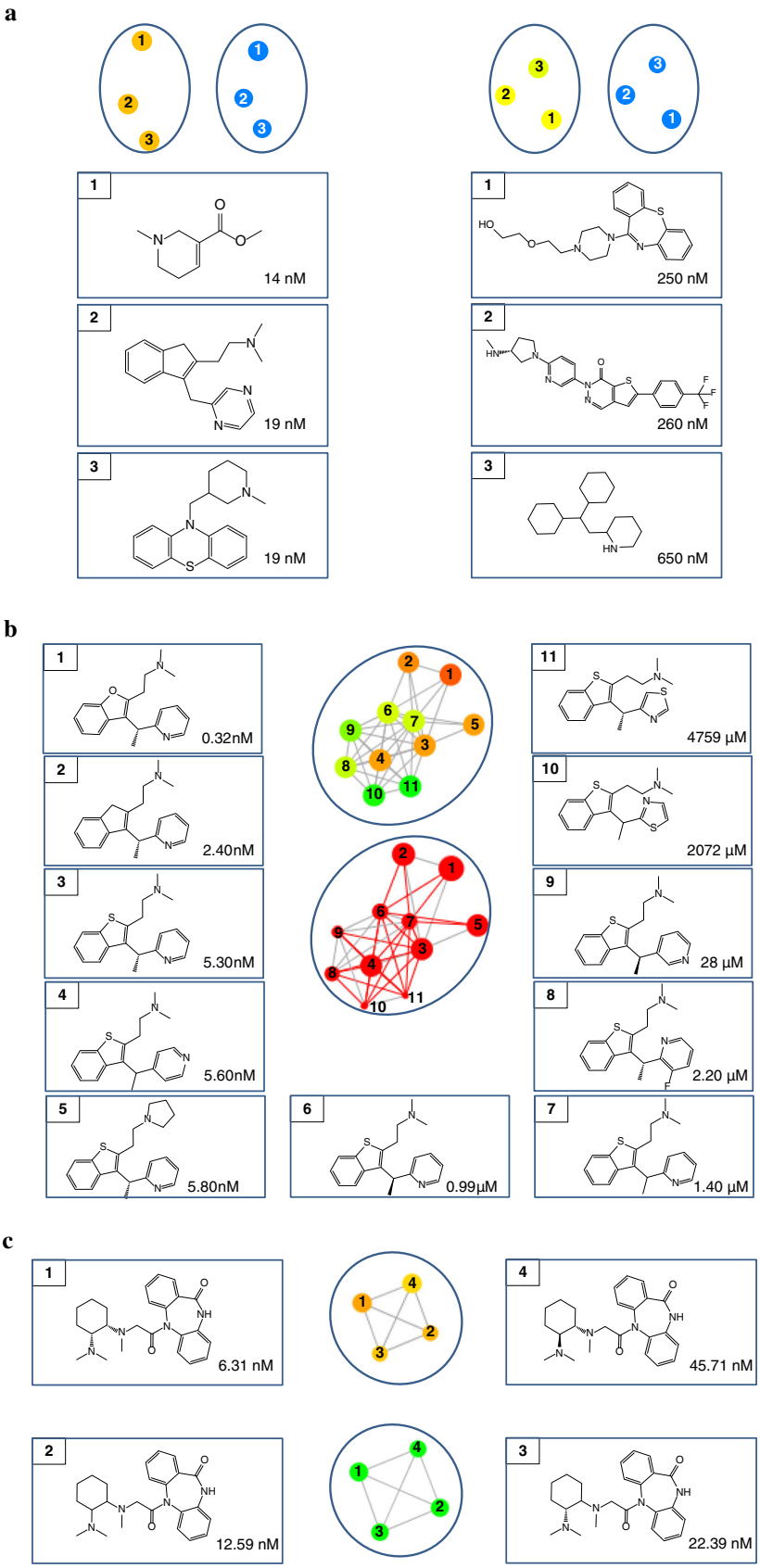


Fig. 3 Local SAR environments. Compound subsets and clusters *highlighted* in Fig. 2a are shown including **a** subsets 1 and 2 representing compounds likely to form similarity cliffs, **b** cluster 3 with compounds likely to form activity cliffs, and **c** cluster 4 with compounds likely to form smooth pairs. Compounds are numbered in the corresponding subgraphs taken from the NSG and FP_NSG and their structures and potency values are reported

compounds forming different AL features. Moreover, they enable the differentiation of compound data sets at a global level. For example, the NSG representations of the data sets in Fig. 2a, b are qualitatively rather similar. However, the corresponding FP_NSGs reveal that similarity cliffs are prevalent in the compound set in Fig. 2a, whereas the set in Fig. 2b is rich in compounds that cannot be assigned to defined AL feature categories. Thus, despite the presence of similar NSGs, these two compound sets are characterized by different global SAR information content. Furthermore, in NSGs, compounds that do not form similarity relationships to others represent singletons that cannot be further considered in SAR analysis. By contrast, in FP_NSGs, singletons are further differentiated. They are either unassigned (and hence not informative in the context of our SAR network analysis) or have a high probability to form similarity cliffs, as also revealed in Fig. 2b, c. The subset of similarity cliff compounds then delineates the major scaffold hopping region within a data set. This information can principally not be obtained from NSGs. Moreover, FP_NSGs further elucidate relationships between global and local SAR features. In NSGs, many local SAR environments become apparent that indicate a certain degree of discontinuity, as discussed above. However, FP_NSGs then identify those environments that contain the majority of compounds likely to form activity cliffs, as is well illustrated by comparing clusters 1-3 side-by-side in Fig. 2b and also highlighted in Fig. 3b. Thus, taken together, the findings reported herein reveal a clear gain in SAR information, both at the global and local level, by taking FP_NSGs into consideration.

Conclusions

In this work, we have reported the design of a network-based AL variant to visualize probabilities of individual compounds to form activity cliffs, similarity cliffs, and smooth pairs, which represent primary features of ALs. Taking per-compound probabilities into account further refines the evaluation of AL features that are typically assessed at the level of compound pairs, rather than individual molecules. Visualization of feature probability distributions in data sets further increases AL information content. The NSG representation, an original network-based AL design, and the FP_NSG representation, as introduced herein, provide complementary SAR information at a global and local level. The FP_NSG representation makes it possible to further distinguish between different compound subsets/clusters and identify key compounds that are most likely to form important features such as activity cliffs within a given data set. When data sets grow in size, SAR information associated with such compounds might be preferentially monitored to further evaluate their role as potential SAR determinants and focal points of optimization efforts. The FP_NSG implementation is freely available to academic investigators upon request.

Acknowledgments The authors thank Dr. Dilyana Dimova for providing activity landscape models using alternative similarity measures and Dr. Ye Hu for help with data sets. B.Z. is supported by the China Scholarship Council.

References

1. Wassermann AM, Wawer M, Bajorath J (2010) Activity landscape representations for structure–activity relationship analysis. J Med Chem 53(23):8209–8223
2. Stumpfe D, Bajorath J (2012) Methods for SAR visualization. RSC Adv 2(2):369–378
3. Iyer P, Stumpfe D, Vogt M, Bajorath J, Maggiora GM (2013) Activity landscapes, information theory, and structure–activity relationships. Mol Inf 32(5–6):421–430
4. Wawer M, Peltason L, Weskamp N, Teckentrup A, Bajorath J (2008) Structure–activity relationship anatomy by network-like similarity graphs and local structure–activity relationship indices. J Med Chem 51(19):6075–6084
5. Peltason L, Iyer P, Bajorath J (2010) Rationalizing three-dimensional activity landscapes and the influence of molecular representations on landscape topology and the formation of activity cliffs. J Chem Inf Model 50(6):1021–1033
6. Stumpfe D, Bajorath J (2012) Exploring activity cliffs in medicinal chemistry. J Med Chem 55(7):2932–2942
7. Stumpfe D, Hu Y, Dimova D, Bajorath J (2014) Recent progress in understanding activity cliffs and their utility in medicinal chemistry. J Med Chem 57(1):18–28
8. Vogt M, Iyer P, Maggiora GM, Bajorath J (2013) Conditional probabilities of activity landscape features for individual compounds. J Chem Inf Model 53(7):1602–1612
9. Willett P, Barnard JM, Downs GM (1998) Chemical similarity searching. J Chem Inf Comput Sci 38(6):983–996
10. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50(5):742–754
11. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP (2012) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40(Database issue):D1100–D1107
12. OEChem TKv 2013 April (2013) OpenEye Scientific Software Inc, Santa Fe, New Mexico