JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

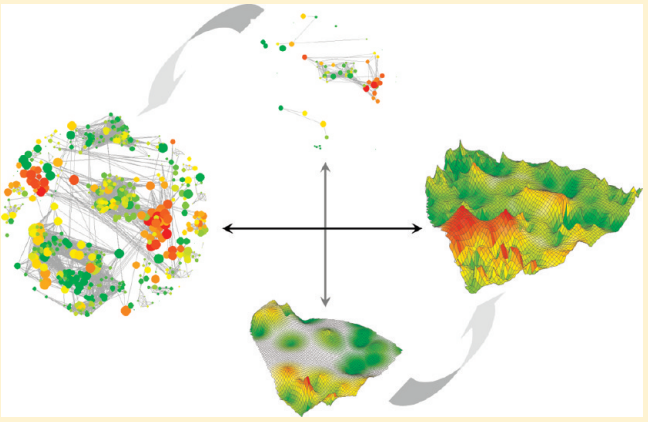
SAR Monitoring of Evolving Compound Data Sets Using Activity Landscapes

Preeti Iyer,[†] Ye Hu,[†] and 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

Supporting Information

ABSTRACT: In pharmaceutical research, collections of active compounds directed against specific therapeutic targets usually evolve over time. Small molecule discovery is an iterative process. New compounds are discovered, alternative compound series explored, some series discontinued, and others prioritized. The design of new compounds usually takes into consideration prior chemical and structure−activity relationship (SAR) knowledge. Hence, historically grown compound collections represent a viable source of chemical and SAR information that might be utilized to retrospectively analyze roadblocks in compound optimization and further guide discovery projects. However, SAR analysis of large and heterogeneous sets of active compounds is also principally complicated. We have subjected evolving compound data sets to SAR monitoring using activity landscape models in order to evaluate how composition and SAR characteristics might change over time. Chemotype and potency distributions in evolving data sets directed against different therapeutic targets were analyzed and alternative activity landscape representations generated at different points in time to monitor the progression of global and local SAR features. Our results show that the evolving data sets studied here have predominantly grown around seed clusters of active compounds that often emerged early on, while other SAR islands remained largely unexplored. Moreover, increasing scaffold diversity in evolving data sets did not necessarily yield new SAR patterns, indicating a rather significant influence of "me-too-ism" (i.e., introducing new chemotypes that are similar to already known ones) on the composition and SAR information content of the data sets.



INTRODUCTION

Structure−activity relationship (SAR) analysis is of central relevance in medicinal chemistry and a major application area for different types of computational methods.[1] Approaches for the qualitative or quantitative analysis of SARs and the prediction of active compounds include classical linear and nonlinear QSAR,[2] pharmacophore,[3] molecular shape,[4] or machine learning[5] methods. SAR exploration in pharmaceutical research makes use of computational approaches in different ways. For example, predictive computational models might be used to aid in optimization of active compounds. Furthermore, data mining methods can be applied to search for and extract available SAR information from large compound data sets. Biological screening data are a prime example of large data sets that are subjected to initial SAR exploration to identify promising hits. For this purpose, different types of graphical analysis tools have been introduced.[6−8] However, screening sets are not the only large compound data source that need to be mined for SAR information. Other data sets often accumulate and grow over time in pharmaceutical environment that contain compounds from different stages of hit-to-lead or lead optimization projects, carried out at different points in time and making use of incrementally advanced information about bioactive compounds. This is typically the case for high-profile therapeutic targets. For example, hundreds or thousands of inhibitors of protein kinases have accumulated over time in different therapeutic areas that have originated from different hit identification and optimization efforts and also include many compounds that have not been further pursued at different stages of the development pipeline. Such historically evolving data sets typically represent a rich source of compound and SAR information that mirror development efforts over time. Accordingly, their retrospective analysis might be attempted to determine new directions for project teams or estimate the chances of given active chemotypes to make it through the pipeline.

For the search for and extraction of SAR information from large compound sets, including both screening data and evolving compound collections, computational activity landscape modeling is a rather attractive approach. In general, activity landscapes

Received: December 22, 2010
Published: February 15, 2011

Table 1. Evolving Compound Data Sets[a]

		scaffolds		carbon skeletons		global SARI scores		
year	compounds	recurrent	total	recurrent	total	continuity	discontinuity	SARI
				adenosine A2a receptor (AA2)				
2000	53	0	27	0	24	0.00	0.95	0.03
2001	67	2	36	3	32	0.01	0.93	0.04
2002	163	3	45	3	40	0.00	0.63	0.18
2003	183	4	52	5	43	0.00	0.61	0.20
2004	224	2	73	5	56	0.05	0.64	0.20
2005	313	2	114	2	84	0.20	0.57	0.31
2006	482	5	180	13	119	0.31	0.23	0.54
2007	596	7	216	11	133	0.33	0.20	0.56
				LCK tyrosine kinase (LCK)				
2000	35	0	22	0	18	0.99	0.12	0.94
2001	41	0	25	0	21	0.99	0.14	0.93
2002	113	2	56	2	38	0.25	0.01	0.62
2003	124	0	63	0	45	0.43	0.01	0.71
2004	131	3	66	3	46	0.39	0.01	0.69
2005	139	0	71	0	51	0.44	0.02	0.71
2006	341	1	165	3	106	0.46	0.56	0.45
2007	369	1	180	2	117	0.49	0.58	0.46
				μ-opioid receptor (MOR)				
2003	31	0	13	0	11	0.04	0.05	0.49
2004	70	0	33	1	26	0.15	0.85	0.15
2005	156	3	59	3	48	0.12	0.78	0.17
2006	310	4	131	4	96	0.11	0.97	0.07
2007	473	5	172	6	121	0.08	0.95	0.07

[a] For the evolving AA2, LCK, and MOR data sets, cumulative compound numbers are given for the monitored time period, and the numbers of corresponding scaffolds and carbon skeletons are reported. "Recurrent" means that scaffolds or carbon skeletons extracted from compounds added during a particular year were already contained in the data sets. In addition, global SARI scores and the underlying continuity and discontinuity scores are reported for the growing compounds data sets on a yearly basis.

can be defined as data set representations that integrate structure and potency relationships between compounds having similar biological activity.[9] A characteristic feature of activity landscape models is that they often combine numerical[10] and graphical SAR analysis functions.[9] Activity landscape representations of different design and sophistication have been introduced including both two-dimensional (2-D)[11] and three-dimensional (3-D)[12] landscape views. It has been shown that activity landscape analysis can reveal both global and local SAR features.[9] Thus, this approach would in principle be also well suited to analyze and compare growing compound data sets at different points in time.

Herein, we report the analysis of evolving compound data sets using numerical and graphical SAR analysis methods, which represents a previously unexplored application of activity landscape modeling. To these ends, we have assembled historically grown, and further growing, compound data sets for selected targets from public domain compound sources and determined how compound composition, structural diversity, and global and local SAR features changed over time. For global and local SAR monitoring, we have utilized the SAR Index (SARI),[10] a numerical SAR analysis function, and alternative activity landscape views, including network-like similarity graphs (NSGs),[11] an advanced 2-D landscape representation, and 3-D landscape models obtained by dimension reduction of chemical references spaces and interpolation of compound potency surfaces.[12]

In our analysis, both conserved and systematically changing SAR features were observed as data sets grew over time. However, we also found that the addition of new scaffolds to evolving data sets did often not alter local SAR characteristics and contributed only little additional SAR information. However, local SAR regions were also identified that remained largely unexplored as data sets further grew in size. Taken together, our results indicate that SAR monitoring in activity landscape models might help to avoid oversampling of landscape regions with well-established SAR patterns in data sets and direct compound optimization efforts to less explored SAR regions.

MATERIALS AND METHODS

Compound Data Sets. Evolving data sets of compounds active against three human targets, including the adenosine A2a receptor antagonists (AA2), the LCK tyrosine kinase inhibitors (LCK), and the μ-opioid receptor ligands (MOR), were assembled from the ChEMBL database.[13] The composition of these data sets is summarized in Table 1. In the MOR set, many structurally similar compounds were reported to have mixed (partial) agonist/antagonist activity, in addition to designated antagonists

dx.doi.org/10.1021/ci100505m |J. Chem. Inf. Model. 2011, 51, 532–540

and agonists. This set was included in our analysis as an example for the analysis of evolving data sets containing overlapping activities. Whenever available, K_i values were used as compound potency annotations, and when more than one K_i value was available, the geometric mean was calculated to yield the final potency annotation of a compound. From all selected compounds, heteroatom frameworks (scaffolds)[14] were extracted, and these scaffolds were further converted into carbon skeletons by replacing all heteroatoms to carbon and setting all bond orders to one. Scaffolds and corresponding carbon skeletons were generated using in-house Pipeline Pilot[15] scripts.

SAR Index Scoring. SARI is a numerical SAR analysis function originally introduced to characterize the global SAR phenotype of compound data sets.[10] SARI results from two separately calculated scores: a continuity score and a discontinuity score. The non-normalized (raw) continuity score is calculated as the potency-weighted mean of pairwise compound similarity within a compound set A. The continuity score emphasizes structurally diverse compounds having high potency and small potency differences and hence accounts for SAR continuity

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

$$= \frac{\sum_{((i,j) \in A | i > j)} \left(\mathrm{weight}(i,j) \times \frac{1}{(1 + \mathrm{sim}(i,j))} \right)}{\sum_{((i,j) \in A | i > j)} \mathrm{weight}(i,j)}, \mathrm{weight}(i,j) = \frac{P_i \times P_j}{1 + |P_i - P_j|}$$

P stands for potency and $\mathrm{sim}(i,j)$ for the similarity of compounds i and j calculated here as Tanimoto similarity[16] for MACCS fingerprint[17] representations.

In addition, the raw discontinuity score is calculated as the average pairwise potency difference between compounds multiplied by pairwise similarity

$$\mathrm{disc}_{\mathrm{raw}}(A) = \underset{\left(\begin{smallmatrix} (i,j) \in A | \mathrm{sim}(i,j) > T, \\ |P_i - P_j| > 1, i > j \end{smallmatrix} \right)}{\text{mean}} (|P_i - P_j| \times \mathrm{sim}(i,j))$$

The discontinuity score emphasizes structurally similar compounds with large potency differences and thus accounts for SAR discontinuity and activity cliffs. Therefore, only pairs of compounds with MACCS Tanimoto similarity of greater than 0.65 (similarity threshold T) and at least 1 order of magnitude difference in potency are taken into account for discontinuity score calculations.

The raw continuity and discontinuity scores are converted into Z-scores utilizing the score distribution of a reference panel of compound sets.[10] Z-scores are mapped onto the value range $[0,1]$ by calculating the cumulative probability for each Z-score assuming a normal distribution. The normalized continuity and discontinuity scores are combined to produce the SARI value

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

On the basis of SAR scoring, the global SAR phenotype of a compound data set is assigned to one of three categories, including a predominantly continuous SAR (high SARI scores close to 1), discontinuous SAR (low scores close to 0), or heterogeneous SAR (intermediate scores around 0.5).

The SARI discontinuity score has also been adapted to assess the contribution of individual compounds to SAR discontinuity, which results in a local score. Following this scoring scheme, the discontinuity score is calculated by comparing a compound to all other molecules that are more similar to it than the predefined Tanimoto similarity threshold T. The scores are then normalized by using the individual scores of all compounds in the data set as a reference for Z-score calculations.[11] This per-compound discontinuity score is utilized for NSG generation, as described below.

Activity Landscape Models. *Two-Dimensional Network-Like Similarity Graphs.* A NSG[11] represents similarity and potency relationships in a compound data set as an annotated graph. Nodes represent individual molecules that are connected by edges if their calculated similarity (here MACCS Tanimoto similarity) exceeds a predefined threshold value. For activity landscape display, this threshold was set to 0.80, which yielded a balanced edge density across the different data sets. Nodes are color-coded by potency using a continuous gradient from green (lowest potency in the data set) via yellow to red (highest potency). The lower and upper boundary of the gradient were set to the highest minimal and lowest maximal potency values in the three data sets, respectively, to ensure that NSGs could be directly compared across different data sets. Nodes were scaled in size according to the SARI per-compound discontinuity score. Large nodes indicate that the potency of a compound significantly differs from its structural neighbors, and combinations of large red and green nodes mark activity cliffs, i.e., very similar compounds having large potency differences,[9,18] representing the extreme form of SAR discontinuity. A graphical layout algorithm is applied that places multiple densely interconnected compounds closely together and separates weakly or unconnected groups of compounds from each other.[19] NSGs were drawn using the "igraph"[20] package of R.[21]

Three-Dimensional Landscape Views. For the generation of 3-D activity landscape models,[12] pairwise Euclidean MACCS fingerprint distances were calculated for all data set compounds, and the corresponding coordinate-free fingerprint space was reduced to two dimensions by nonmetric multidimensional scaling[22] as a dimension reduction technique. Compound coordinates were normalized to range from 0 to the maximum observed distance in our data sets such that the range of the planar coordinates and the size of the landscapes corresponded to the overall dissimilarity distribution. From compound potency values, a coherent potency surface was interpolated using the Kriging function[23] as implemented in the "fields" package[24] of R and added as the third dimension to the 2-D fingerprint space projection. The potency axis was scaled for the data sets to range from the overall lowest to highest potency value. The interpolated potency surfaces were color-coded using a gradient from green (lowest potency) to red (highest potency), and the value range mapped to the color gradient was determined by the highest minimal and lowest maximal interpolated potency values across the three data sets. The landscape transparency was scaled according to the density of experimental potency values. Hence, grid points proximal to experimental data points (compound potency values) were opaque, and grid points furthest away from these data points were fully transparent (white). All landscape views were displayed using the same reference perspective.

Consistent Representation. For the NSGs and 3-D activity landscapes, the layout and 2-D space projection, respectively, were generated once for largest cumulative data set. Then NSG and 3-D activity landscapes were generated for each year by iterative addition of the corresponding compounds and their potency information. This ensured that the position of molecules in both visualizations remained constant over the years and that

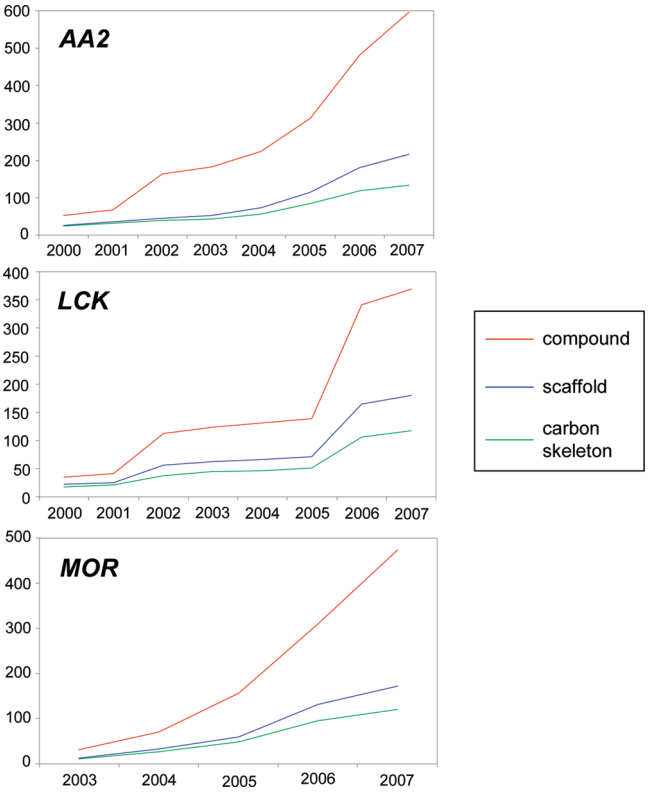


Figure 1. Compound and scaffold distribution. Cumulative numbers of compounds (red), scaffolds (blue), and carbon skeletons (green) are reported over consecutive years for the evolving AA2, LCK, and MOR data sets.

the growth of the data was monitored using a consistent representation frame.

RESULTS AND DISCUSSION

Organization of Evolving Compound Data Sets. We have assembled evolving data sets by collecting specifically active compounds that have been published in subsequent years since 2000. Hence, each data set reflects the progress made in addressing a specific target over a period of time. Active compounds were selected according to publication dates in yearly increments spanning periods of five (MOR) or eight years (AA2 and LCK), as reported in Table 1. Initial sets of approximately 30−50 compounds were obtained that grew over time in size to 369 (LCK), 473 (MOR), or 596 (AA2) compounds.

Chemotype Distribution. In Table 1 and Figure 1, the compound, scaffold, and carbon skeleton distributions are reported for the data sets over time. Carbon skeletons represent topologically distinct scaffolds. As expected, compound numbers increased more rapidly than scaffolds and skeleton numbers. For all three data sets, significant growth occurred after 2005 (Figure 1). Beginning at that time, the gap between compounds and scaffolds substantially widened, indicating that many analogs were added. The curves of the LCK distributions in Figure 1 display very similar traces suggesting that there has been a strong correlation between newly introduced chemotypes and compound numbers. In Table 1, it is also shown that the number of recurrent scaffolds and carbon skeletons per year (i.e., scaffolds/skeletons that were already introduced in previous years) was generally small for all three data sets but especially for LCK. However, for AA2 and MOR, the number of recurrent scaffolds and skeletons notably increased in 2006 and 2007 when compound numbers also grew significantly, consistent with the data distributions in Figure 1. A total of only 16, 6, and 10 scaffolds were recurrent over different years for the AA2, LCK, and MOR sets, respectively, which represented 189, 19, and 75 compounds. Thus, for AA2, recurrent scaffolds ultimately represented ~31% of the compound data set. The AA2, LCK, and MOR compound sets finally contained a total number of 216, 180, and 172 distinct scaffolds, respectively.

Potency Distribution. In Figure 2a, the cumulative compound potency distribution in the three data sets is monitored. For AA2, the median potency essentially remained constant over the years until a slight decrease occurred in 2006 and 2007. For LCK, a notable increase in compound potency was detected in 2002, and the distribution remained constant until 2006 when new highly potent compounds were reported. In the case of MOR, there was a steady decline in median compound potency, although highly potent compounds were added during later years. Thus, increasing numbers of weakly potent compounds became available over the years. Because the number of recurrent scaffolds and carbon skeletons was generally small for all data sets, many of these weakly potent compounds represented new scaffolds. In Figure 2b, non-cumulative potency distributions are shown.

Global SAR Character. In Table 1, SARI scores calculated for the growing data sets are also reported. Here these scores provide a measure for the global SAR phenotype of a data set monitored over time. Globally continuous SARs are characterized by the presence of structurally increasingly diverse active compounds, discontinuous SARs by the presence of many structurally similar compounds with significantly different potency, i.e., activity cliffs, and heterogeneous SARs by the coexistence of multiple continuous and discontinuous SAR components.

For our three compound sets, interesting trends were observed. Two of the data sets, AA2 and LCK, changed their global SAR character in a well-defined manner over time, whereas the SAR phenotype of the MOR set remained constant. In the latter case, after the first year (when only 31 active compounds were available), the MOR data set was characterized by strong global SAR discontinuity (low continuity, high discontinuity, and low global SARI score) that remained largely constant as the compound set significantly grew in size between 2004 and 2007.

During the first two years, the AA2 set was also characterized by strong global discontinuity. Then, however, the discontinuity decreased and continuous SAR components began to evolve, which ultimately resulted in a globally heterogeneous SAR phenotype (SARI score ~0.5) of the final compound set. By contrast, an opposite trend was observed for the LCK data set. During the early stages, this set was strongly continuous in nature. As it further evolved, the SAR continuity decreased and notable discontinuity was observed beginning in 2006, which again resulted in a globally heterogeneous SAR phenotype. Thus, taken together, the evolution of these data sets was accompanied by different changes in global SAR character.

Activity Landscapes. For our compound data sets, 2-D and 3-D landscape representations were generated to monitor how compound similarity and potency relationships in these data sets changed over time and how global and local SAR features evolved. The 3-D landscape representations provide a rather global and intuitive view of an activity landscape. In addition, NSGs are designed to study relationships between global and local SAR features and identify key compounds, for example, molecules forming prominent activity cliffs in a data set. The similarity

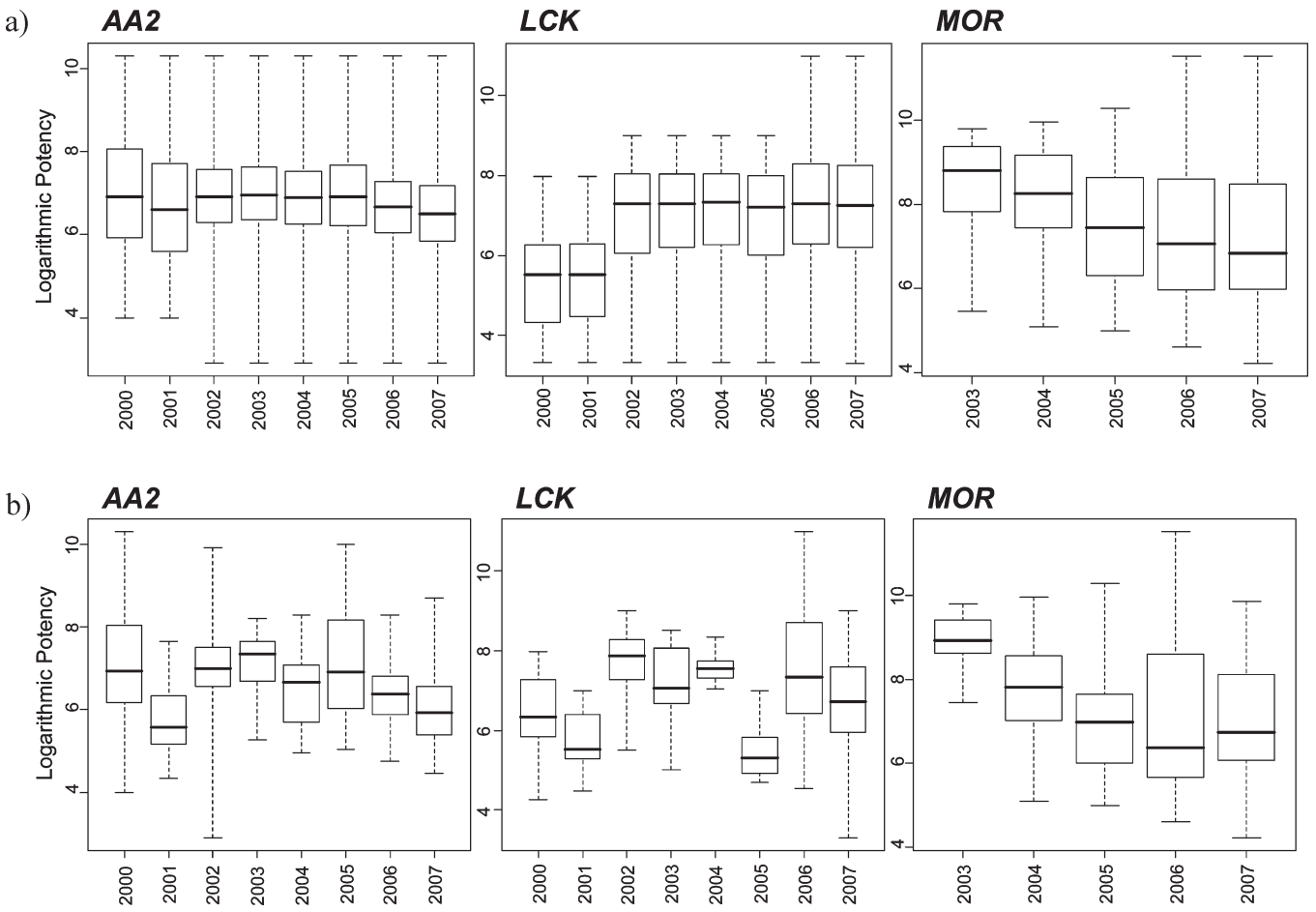


Figure 2. Potency distribution. (a) Box plots report the cumulative distribution of logarithmic potency values for the evolving AA2, LCK, and MOR data sets. Box plots represent value distributions starting from the smallest observed value (bottom) to the lower quartile, median (thick horizontal line), upper quartile, and the largest value (top). (b) Corresponding noncumulative potency distributions, i.e., potency distributions of compound subsets added during each year.

cutoff for edges in these NSGs was chosen such that compounds connected by an edge were visibly similar in structure. In the following, both 2-D and 3-D activity landscape views are provided for those years during the evolution of these data sets when notable changes occurred.

AA2. Panel (a) of Figure 3 shows activity landscapes for the AA2 data set. The 3-D representations illustrate how the activity landscape gradually "filled in" with compounds ("empty" regions are white) and how discontinuous and more continuous landscape regions shaped up. The corresponding NSGs provide further details. The NSG of the 2003 version of the compound set is dominated by a central network component of structurally similar highly and weakly potent compounds that make different contributions to local SAR discontinuity (indicated by different node sizes). Exemplary compounds from this region are depicted in panel (b) of Figure 3. The further growth of the data set was largely centered on four seed areas (1−4 in Figure 3a). Populating these areas with compounds resulted in densely connected clusters that introduced predominantly continuous (1, 3), heterogeneous (2), or discontinuous (4) SAR components into the data set. However, in regions 1−3, the addition of significant numbers of new compounds over time did not yield significant amounts of new SAR information because the SAR character of these NSG regions did not notably change over time. Thus, these regions might be deprioritized for further compound and SAR exploration. However, in region 4, new compounds introduced a number of prominent activity cliffs, and this region would provide a more promising focal point for further exploratory efforts. All AA2 activity landscapes display increasing heterogeneity over time, consistent with the global SAR phenotype.

LCK. The 2-D and 3-D activity landscapes for the years 2002 and 2005 of LCK in panel (a) of Figure 4 represent data sets of similar size including the presence of only a limited number of potent compounds that were characterized by global SAR continuity. Between 2000 and 2005, these characteristics remained fairly constant. However, in 2006, the LCK compound set more than doubled in size, and a number of highly potent compounds became available that predominantly formed four separate clusters (1−4 in Figure 4a). In addition, two of these regions (1, 2) and several other less populated clusters also contained structurally very similar compounds with large potency differences, thus producing overall distinctly heterogeneous activity landscapes, which becomes clearly apparent in the 2-D and 3-D landscape representations for 2006 and 2007. Given their compound composition and SAR character, region 1 and especially region 2 would be attractive candidates for further SAR exploration of this data set. Panel (b) of Figure 4 shows compounds forming exemplary activity cliffs found in region 2. However, potent compounds did

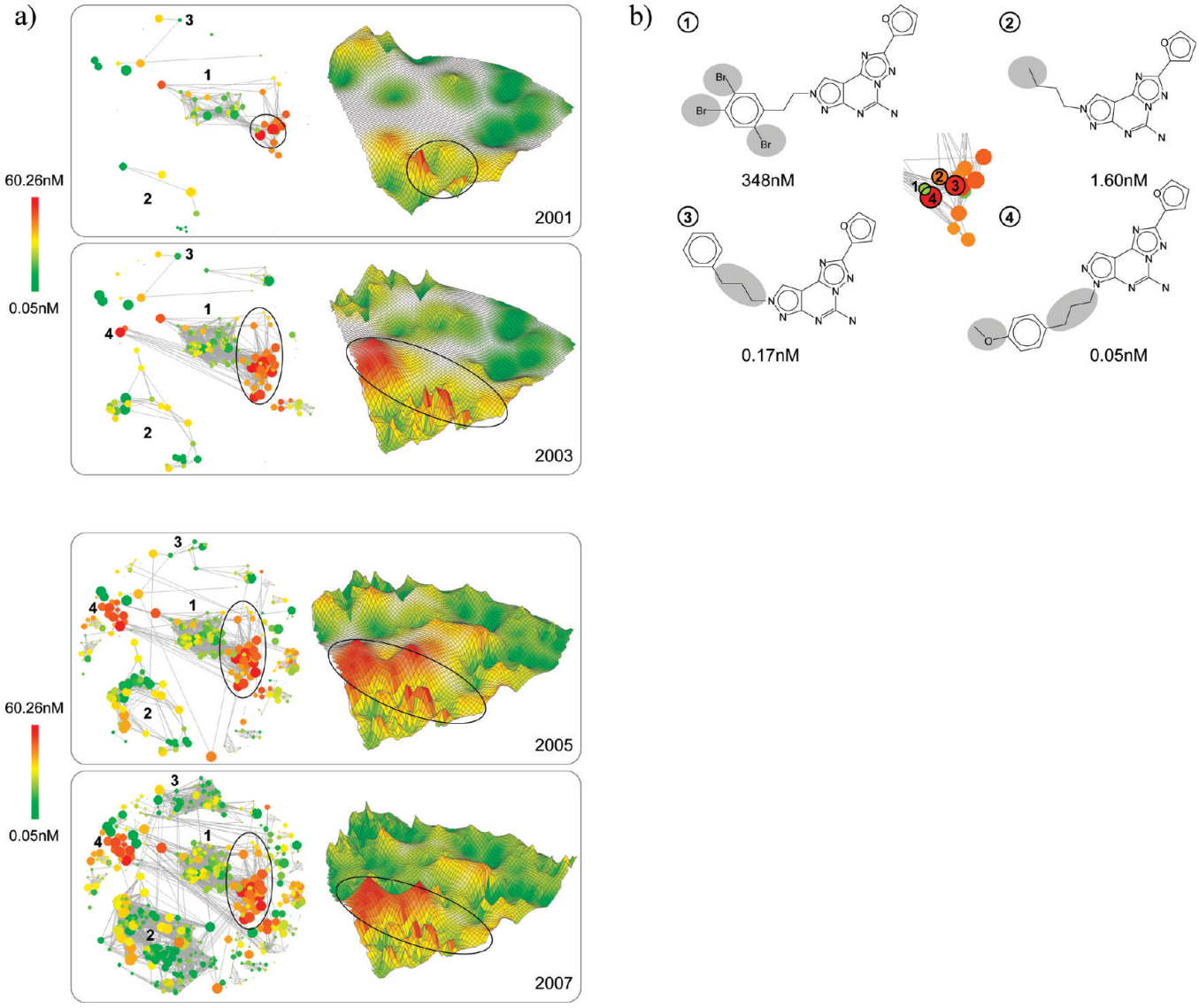


Figure 3. Activity landscape representations of the AA2 data set. (a) NSGs and 3-D activity landscapes are shown for the years 2001, 2003, 2005, and 2007. These activity landscapes are representative of the evolution of the data set between 2000 and 2007. The activity landscapes of in-between years that are not displayed are very similar to one of the landscapes shown here because of the limited numbers of compounds added during these intermittent years (see also Table 1). However, activity landscape representations for all years are shown in Figure S1 of the Supporting Information. Groups of compounds labeled with numbers (1−4) indicate NSG regions that substantially grow over time and/or change in SAR character. A group of compounds for which structures are shown in (b) is encircled in both 2-D and 3-D landscape views. For landscape display, the potency color code was adjusted to the potency range covered by all three data sets (shown on the left), which provided a constant node color scheme for all activity landscape representations. (b) For a selected NSG region (in the 2001 graph) encircled in (a), structures of compounds making different contributions to SAR discontinuity are shown. Node and structure labels correspond to each other, and structural differences between these compounds are shown on a gray background.

not always introduce SAR discontinuity into the LCK set. For example, regions 3 and 4 consist of structurally very similar compounds having comparably high potency that do not have structural neighbors with lower potency. Hence, these compounds are related to each other by a continuous local SAR and provide only little new SAR information. Overall, the activity landscapes reflecting the evolution of the LCK compound set also illustrate the change in global data set characteristics, consistent with the conclusions drawn from numerical SARI profiling.

MOR. The MOR data set differed from the other two compound sets on the basis of SAR profiling because it was characterized by the presence of strong SAR discontinuity that did not notably change over time. The activity landscapes of the evolving MOR data set shown in panel (a) of Figure 5 corroborate this finding. The activity landscapes contain many pronounced activity cliffs. The dominance of activity cliffs becomes especially apparent in the 3-D landscape representations for 2006 and 2007. In NSGs, prominent activity cliffs are marked by combinations of large red and green nodes. However, despite the overall high degree of discontinuity, the activity landscapes also contain regions of different local SAR character. In panel (a) of Figure 5, six NSGs regions are labeled (1−6) that have grown at different rates and represent different local SAR environments. For example, region 1 represents an island of strong local SAR

dx.doi.org/10.1021/ci100505m |*J. Chem. Inf. Model.* 2011, 51, 532–540

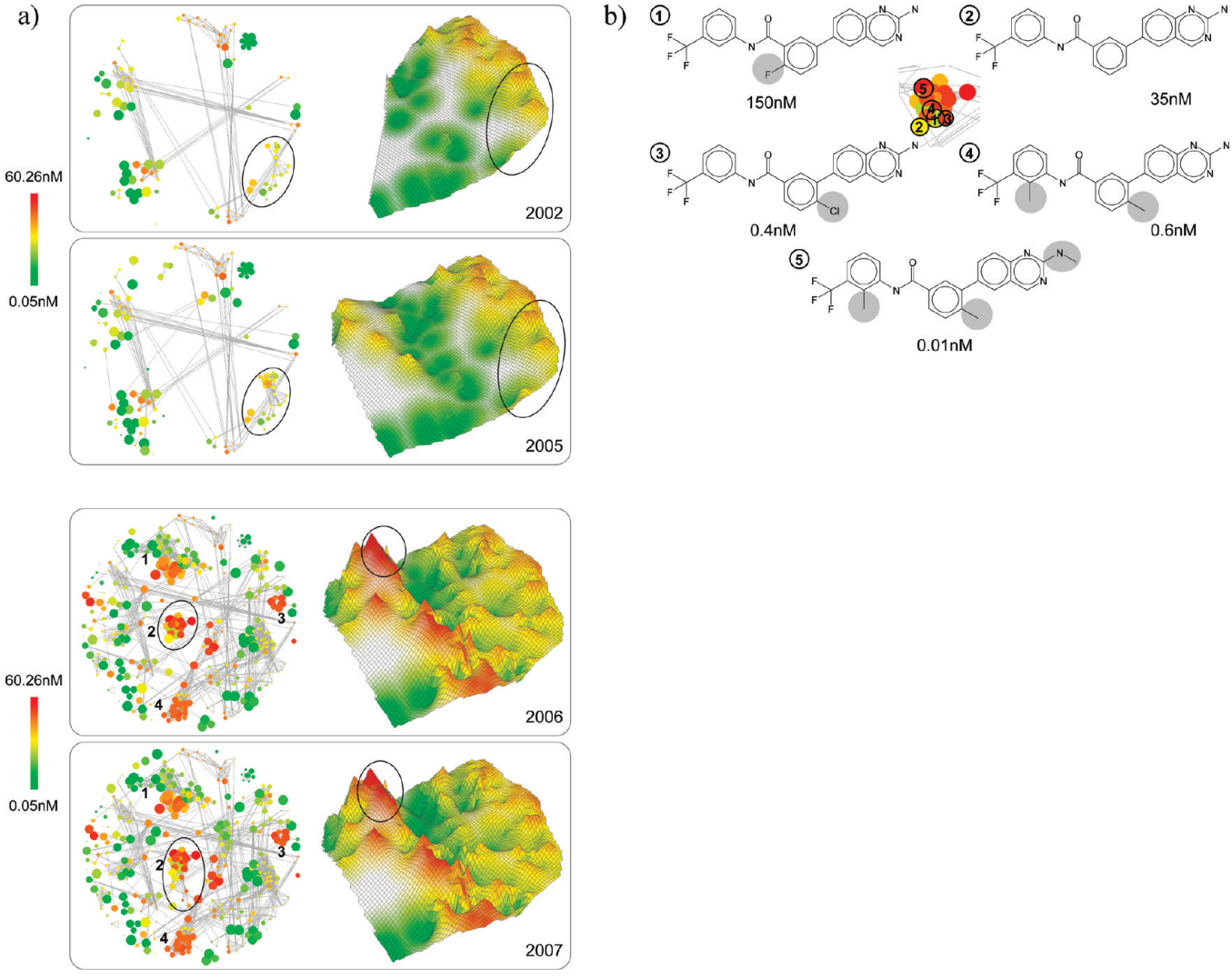


Figure 4. Activity landscape representations of the LCK data set. (a) NSGs and 3-D activity landscapes are shown for the years 2002, 2005, 2006, and 2007. These activity landscapes are representative of the evolution of the data set between 2000 and 2007. Activity landscape representations for all years are shown in Figure S2 of the Supporting Information. Groups of compounds that represent selected local SAR environments are labeled (1−4), and a group of compounds forming exemplary activity cliffs is encircled in both 2-D and 3-D landscape views. (b) For a selected NSG region (in the 2006 graph) encircled in (a), structures of compounds forming activity cliffs are shown. Node and structure labels correspond to each other, and differences between these compounds are shown on a gray background.

continuity that did not significantly change over time. This region contains structurally very similar and moderately potent compounds. By contrast, from the beginning, region 2 has been a center of activity cliff formation. Exemplary compounds from this region are shown in panel (b) of Figure 5. Over several years, no similar compounds were reported until in 2007 a large number of weakly and moderately potent compounds were added to this region. However, many of these compounds make strong contributions to SAR discontinuity because they form prominent activity cliffs with highly potent compounds in region 5, which has been another center of growth for this data set. This region contains many potent analogs of four potent compounds that first appeared in 2005. The relationship between regions 2 and 5 is a major source of SAR discontinuity and points at a characteristic feature of this data set. It contains groups of structurally highly similar compounds (densely connected clusters in NSGs) that are either weakly or highly potent and form a multitude of activity cliffs across clusters. Thus, given the dense sampling of these regions, a substantial amount of redundant compound structure information has been generated over the years. This type of relationship is also seen for regions 3, 2, and 4, where many small to moderately sized cliffs are formed across clusters of very similar compounds. Thus, the activity landscape representations reveal many local SAR environments of varying character in this globally strongly discontinuous data set. For compound exploration, further a sampling of regions 2 and 5 would be less attractive than focusing, for example, on region 4 that has thus far been much less explored but contains compounds making rather different contributions. Furthermore, it would also not be very attractive to further study region 1 because the compounds forming this region are related to each other by a "flat" local SAR. By contrast, region 6 would be another attractive area for further SAR analysis, given its distinct heterogeneity and the relatively low density of structurally similar compounds.

Common Features. Although the activity landscapes and SAR characteristics of the compound data sets studied here differed in

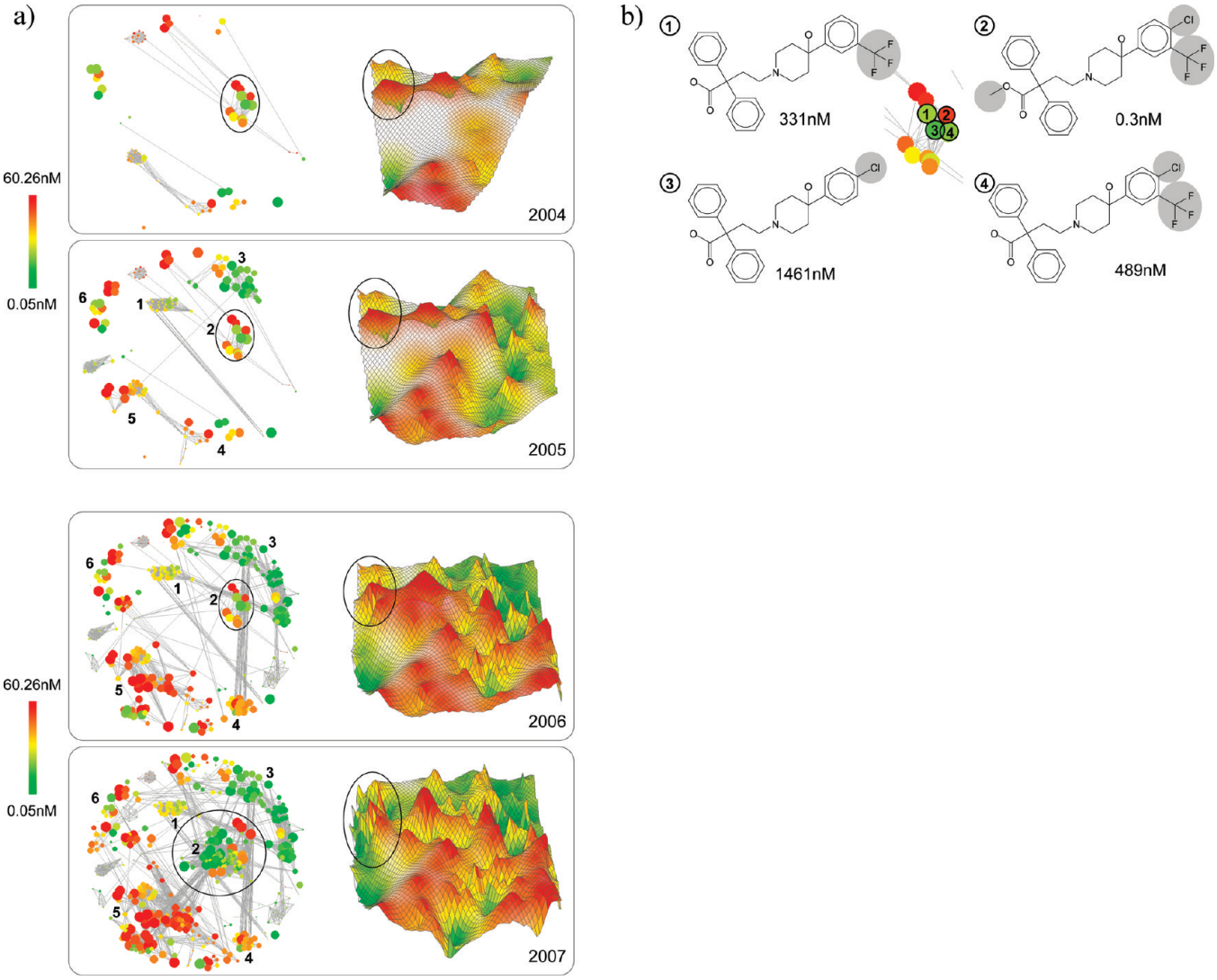


Figure 5. Activity landscape representations of the MOR data set. (a) NSGs and 3-D activity landscapes are shown for the years 2004, 2005, 2006, and 2007. These activity landscapes are representative of the evolution of the data set between 2003 and 2007. Activity landscape representations for all years are shown in Figure S3 of the Supporting Information. Groups of compounds that represent selected local SAR environments are labeled (1−6), and a group of compounds forming exemplary activity cliffs is encircled in both 2-D and 3-D landscape views. (b) For a selected NSG region (in the 2004 graph) encircled in (a), structures of compounds forming activity cliffs are shown. Node and structure labels correspond to each other, and differences between these compounds are shown on a gray background.

part significantly, some common features were identified. For example, data set growth was usually centered on a few different compound subsets (local landscape regions), but other compound subsets that appeared over time were not further explored. In addition, the landscape representations of all data sets displayed a substantial degree of SAR heterogeneity. However, compound clusters were also observed that grew substantially without notable changes in local SAR character. Ultimately, these clusters contained redundant compound structure information. This observation leads to a rather important point revealed by our analysis. As discussed above, only small numbers of recurrent scaffolds and carbon skeletons were generally observed and, in addition, the compound-to-scaffold ratios were rather low. These findings would indicate that many novel compounds were generated over the years. However, activity landscape analysis revised this view because in all cases many densely connected clusters consisting of very similar compounds were observed. Thus, many "novel" scaffolds were apparently obtained by making only small modifications to previously reported chemotypes such as, for example, variations of linker length, individual rings, or heteroatom positions. Accordingly, many of the corresponding compounds did not yield new SAR information. However, in all compound data sets, regardless of their global SAR characteristics, we also detected evolving regions that were rich in SAR information due to the presence of many activity cliffs.

CONCLUDING REMARKS

In this study, we have monitored target-centric evolving compound data sets using statistical (hierarchical structure and potency) and activity landscape analysis. In pharmaceutical research, the study of evolving data sets is highly relevant in order to rationalize and guide compound exploration and optimization efforts in the course of long-term discovery projects. SAR monitoring

of the literature data sets we assembled revealed some systematic trends in the progression of global and local SAR characteristics. Conclusions drawn on the basis of numerical SAR profiling and graphical activity landscape analysis were overall consistent. NSG analysis would essentially be sufficient to draw major conclusions concerning the evolving data sets studied here. However, we have deliberately utilized 2-D and 3-D landscape representations, which fundamentally differ in their design and details. The alternative landscape views revealed evolving landscape feature in a graphically complementary manner. The 3-D landscape models provide a global view of a data set, and NSGs are designed to reveal relationships between global and local SAR features. They also provide a global data representation but focus the analysis on local SAR environments. SAR monitoring using these alternative activity landscape views is essentially not limited by compound set size or potency distributions. Both NSGs and 3-D landscape models can be generated for large data sets, and color coding of potency distributions can be easily adjusted.

The analysis of activity landscapes generated for data sets at different time points during their evolution revealed a general heterogeneity of global and local SAR features and identified regions on which the growth of these data sets was centered. In addition, we found that structural novelty among active compounds was usually more limited than indicated on the basis of molecular scaffold analysis and that many newly introduced compounds did not provide significantly more SAR information than already available because they complemented existing densely populated clusters with well-defined SAR patterns. For SAR monitoring and the retrospective analysis of evolving compound data sets, the identification of such oversampled SAR regions is an important aspect. Such findings should help to redirect compound exploration efforts from compound subsets with well-defined SAR behavior to regions in activity landscapes that are not well sampled and where the addition of compounds might yield more SAR information and help to identify new leads.

■ ASSOCIATED CONTENT

Supporting Information. Figures S1−S3 show activity landscape representations for all monitored years and the AA2, LCK, and MOR data sets, respectively. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Tel: +49-228-2699-306; fax: +49-228-2699-341; e-mail: bajorath@bit.uni-bonn.de.

Author Contributions

†The contributions of these authors should be considered equal.

■ REFERENCES

(1) Peltason, L.; Bajorath, J. Systematic computational analysis of structure−activity relationships: Concepts, challenges and recent advances. *Future Med. Chem.* **2009**, *1*, 451–466.

(2) Esposito, E. X.; Hopfinger, A. J.; Madura, J. D. Methods for applying the quantitative structure−activity relationship paradigm. *Methods Mol. Biol.* **2004**, *275*, 131–214.

(3) Leach, A. R.; Gillet, V. J.; Lewis, R. A.; Taylor, R Three-dimensional pharmacophore methods in drug discovery. *J. Med. Chem.* **2010**, *53*, 539–558.

(4) Nicholls, A.; McGaughey, G. B.; Sheridan, R. P.; Good, A. C.; Warren, G.; Mathieu, M.; Muchmore, S. W.; Brown, S. P.; Grant, J. A.; Haigh, J. A.; Nevins, N.; Jain, A. N.; Kelley, B. Molecular shape and medicinal chemistry: A perspective. *J. Med. Chem.* **2010**, *53*, 3862–3886.

(5) Geppert, H.; Vogt, M.; Bajorath, J. Current trends in ligand-based virtual screening: Molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205–216.

(6) Ahlberg, C. Visual exploration of HTS databases: Bridging the gap between chemistry and biology. *Drug Discovery Today* **1999**, *4*, 370–376.

(7) Roberts, G.; Myatt, G. J.; Johnson, W. P.; Cross, K. P.; Blower, P. E., Jr. LeadScope: Software for exploring large sets of screening data. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1302–1314.

(8) Wawer, M.; Peltason, L.; Bajorath, J. Elucidation of structure−activity relationship pathways in biological screening data. *J. Med. Chem.* **2009**, *52*, 1075–1080.

(9) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity landscape representations for structure−activity relationship analysis. *J. Med. Chem.* **2010**, *53*, 8209–8223.

(10) Peltason, L.; Bajorath, J. SAR Index: Quantifying the nature of structure−activity relationships. *J. Med. Chem.* **2007**, *50*, 5571–5578.

(11) 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.

(12) Peltason, L.; Iyer, P.; Bajorath, J. Rationalizing three-dimensional activity landscapes and the influence of molecular representations on landscape topology and the formation of activity cliffs. *J. Chem. Inf. Model.* **2010**, *50*, 1021–1033.

(13) *ChEMBL*; European Bioinformatics Institute (EBI): Cambridge, 2010. http://www.ebi.ac.uk/chembl/ (accessed November 02, 2010).

(14) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(15) *SciTegic's Pipeline Pilot*, Student ed., version 6.1; Accelrys, Inc.: San Diego, 2007.

(16) Willett, P. Searching techniques for databases of two- and three-dimensional structures. *J. Med. Chem.* **2005**, *48*, 1–17.

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

(18) Maggiora, G. M. On outliers and activity cliffs: Why QSAR often disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535–1535.

(19) Fruchterman, T. M. J.; Reingold, E. M. Graph drawing by force-directed placement. *Software—Pract. Exper.* **1991**, *21*, 1129–1164.

(20) Csardi, G. *The igraph library*, version 0.5.5; Budapest, Hungary, 2009. http://igraph.sourceforge.net/ (accessed December 16, 2010).

(21) *R: A Language and Environment for Statistical Computing*; R Development Core Team, R Foundation for Statistical Computing: Vienna, Austria, 2010.

(22) Borg, I.; Groenen, P J. F. *Modern Multidimensional Scaling. Theory and Applications*, 2nd ed.; Springer: New York, 2005.

(23) Cressie, N. *Statistics for Spatial Data*, revised ed.; Wiley: New York, 1993.

(24) Furrer, R.; Nychka, D.; Sain, S. *Tools for Spatial Data, R Package*, version 6.30; R Foundation for Statistical Computing: Vienna, Austria, 2009.