JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

ARTICLE

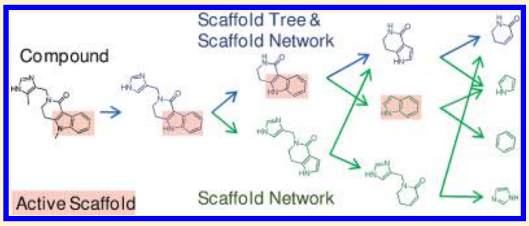
pubs.acs.org/jcim

Mining for Bioactive Scaffolds with Scaffold Networks: Improved Compound Set Enrichment from Primary Screening Data

Thibault Varin, Ansgar Schuffenhauer, Peter Ertl, and Steffen Renner*

Novartis Institutes for BioMedical Research, Forum 1, Novartis Campus, CH-4056 Basel, Switzerland

ABSTRACT: Identification of meaningful chemical patterns in the increasing amounts of high-throughput-generated bioactivity data available today is an increasingly important challenge for successful drug discovery. Herein, we present the scaffold network as a novel approach for mapping and navigation of chemical and biological space. A scaffold network represents the chemical space of a library of molecules consisting of all molecular scaffolds and smaller "parent" scaffolds generated therefrom by the pruning of rings, effectively leading to a network of common scaffold substructure relationships. This algorithm provides an extension of the scaffold tree algorithm that, instead of a network, generates a tree relationship between a heuristically rule-based selected subset of parent scaffolds. The approach was evaluated for the identification of statistically significantly active scaffolds from primary screening data for which the scaffold tree approach has already been shown to be successful. Because of the exhaustive enumeration of smaller scaffolds and the full enumeration of relationships between them, about twice as many statistically significantly active scaffolds were identified compared to the scaffold-tree-based approach. We suggest visualizing scaffold networks as islands of active scaffolds.



INTRODUCTION

With the majority of today's high-throughput drug discovery approaches, one of the main challenges remains the analysis and visualization of very large data sets of molecules with associated biological data, such as a screening library with data from a high-throughput screening (HTS) campaign. One of the main goals of this process is the identification and understanding of chemical patterns responsible for the desired activity. Molecular scaffolds are one of the most appealing concepts in medicinal chemistry to describe, discuss, and visualize series of chemical compounds and associated biological properties in an aggregated manner. The concept is rich in information and still easily understood by chemists, computational chemists, and biologists, and it seems to reflect well the way bioactive molecules are developed by chemists and by natural evolution. In their seminal work, Bemis and Murcko[1] discovered that the majority of drugs analyzed in their study could be represented by only a small set of scaffolds. They defined scaffolds as the framework of the molecular topological graph that is obtained by the removal of all terminal side chains. Various applications of scaffolds have been reported from the assessment of library diversity,[2−5] scaffold hopping,[6,7] or of scaffold target/biological property profiles.[8,9]

In 2005, two similar approaches for the analysis of large chemical data sets, possibly with associated biological properties, were published for the scaffold-based mapping and analysis of chemical space, namely, HierS[10] and the SCONP tree of natural products,[11] which was further developed to the scaffold tree (ST) approach.[12,13] Both approaches are based on the idea of defining a hierarchy of scaffolds by iteratively removing rings from the molecule's scaffold defined by the original Bemis—Murcko idea (henceforth called the Murcko scaffold). With HierS, all combinations of smaller scaffolds that could be generated by the removal of fused ring systems are generated from the Murcko scaffold, essentially leading to a network of scaffolds. In contrast, in SCONP and the scaffold tree approach, only one representative scaffold is retained for each molecule on each level of the scaffold hierarchy, based on scaffold frequency (SCONP) or a sophisticated set of medicinal-chemistry-derived rules designed to identify chemically interesting scaffolds (scaffold tree). Here, smaller scaffolds are generated by the iterative removal of single rings, also dissecting fused ring systems. The SCONP/scaffold tree strategy allows for an elegant tree representation of the library, for which several software visualization tools have since been developed.[14−18] We showed recently that a scaffold-hierarchy-based description of chemical and bioactivity space reflects well the way bioactivity is distributed over compound libraries.[19] Long branches of scaffolds from structurally complex to simple scaffolds with retained but varying bioactivity were found with high frequency for the five major pharmaceutically relevant target classes and allowed for the identification of new inhibitor types for a given target by filling gaps in the scaffold hierarchy. Such analyses have led to a number of chemical libraries synthesized around scaffolds considered to be interesting according to the scaffold tree approach.[18−21]

Based on the scaffold tree concept, we recently proposed the compound set enrichment (CSE) approach,[22] which is able to identify statistically significantly enriched scaffolds from primary screening data. In contrast to traditional methods that consider only the activities of individual molecules, for example, selecting

Received: February 22, 2011
Published: May 26, 2011

dx.doi.org/10.1021/ci2000924 | *J. Chem. Inf. Model.* 2011, 51, 1528–1538

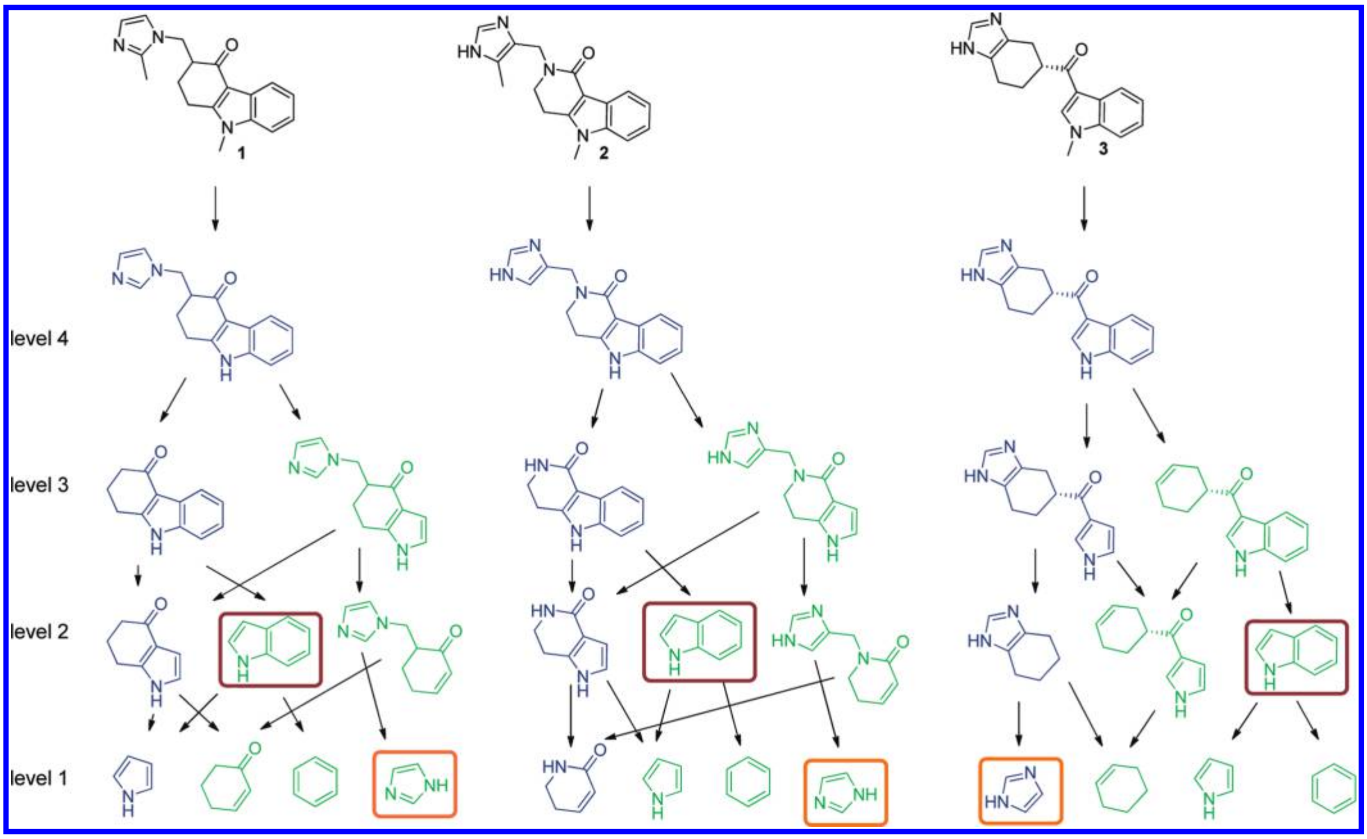


Figure 1. Example of scaffold trees and scaffold networks generated from three 5-HT3 antagonists: Ondasetron (**1**), Alosetron (**2**), and Ramosetron (**3**). Blue scaffolds were generated by both the scaffold tree and scaffold network approaches, whereas green scaffolds were generated uniquely by the scaffold network approach. In this example, indoles (highlighted with brown boxes) would be generated only by the SN decomposition and, therefore, would not be recognized as conserved scaffolds by the scaffold tree approach. For imidazoles (highlighted with orange boxes), one of the three occurrences is detected by the scaffold tree approach (for Ramosetron, **3**); however, the ST approach does not recognize the presence of imidazoles in Ondasetron and Alosetron. Considering all occurrences of a scaffold in a library (i.e., using more samples) enables a better statistical assessment of a scaffolds' activity and might identify weaker signals than using fewer examples.

the *x* most active molecules from a screen or all molecules with activities that satisfy a predefined threshold, statistical scaffold-based methods are able to identify series of molecules that are statistically significantly active in the desired way as an ensemble. As such, activity should be less affected by statistical errors and might enable the identification of "latent hits", scaffolds of weak hits, possibly even fragments that have the potential to be optimized into potent druglike molecules. The CSE method generates all scaffold tree scaffolds from a screening library and calculates for each scaffold *p*-values for a biological measurement (here, primary HTS readouts) using Kolmogorov−Smirnoff (KS) statistics. This approach successfully identified the known active scaffolds as determined by dose−response curves, using only the primary screening readout. In addition, a large number of previously unidentified active scaffolds from PubChem bioassays were identified.

This study used scaffolds generated by the scaffold tree algorithm. Because of the rule-based generation of a scaffold tree retaining only one single scaffold at each level of the tree, we reasoned that a number of possible scaffolds not considered in the latter analysis might be interesting in terms of biological activity. Therefore, we aimed to test the effect of considering all possible permutations of smaller scaffolds for molecules in the CSE approach. This extension of the scaffold tree is termed the "scaffold network" (SN). Scaffold networks are conceptually similar to the HierS algorithm[10] in generating a network of smaller scaffolds rather than a tree, as with the scaffold tree approach. In contrast to HierS, we keep the scaffold tree philosophy of dissecting fused ring systems instead of pruning only entire fused ring systems at a time, as is done in HierS. An example of the conceptual difference between the scaffold tree and scaffold network approaches is given in Figure 1, a hierarchical scaffold decomposition of the three 5-HT3 antagonists Ondasetron (**1**), Alosetron (**2**), and Ramosetron (**3**). The original scaffold tree algorithm generated only a single scaffold per hierarchy level (blue), whereas the scaffold network approach generated all possible scaffolds at each hierarchy level (additional scaffolds shown in green). In this particular example, the scaffold tree algorithm did not detect any common scaffolds between the structurally and pharmacologically related molecules. With the exhaustive scaffold decomposition of the scaffold network approach, indoles and imidazoles suddenly emerged as fully conserved features of the ligands, which is nicely in agreement with the known requirements of 5-HT3 antagonists.[23] Using the scaffold tree algorithm alone would have missed these real relationships.

When this article was in the final stages before submission, the Bajorath group published an article comparing scaffold tree decomposition with the substructure relationships of Murcko scaffolds of molecules that are active toward the same target.[24] In their study, they found that only about 32% of such pairs of scaffolds were identified by scaffold tree decomposition, and they suggested using scaffold substructure relationships as an additional orthogonal means for scaffold tree analysis. We agree with Hu and Bajorath's conclusions that the ST approach might not be sufficient for the analysis of bioactivity, but we believe that scaffold substructure relationships still miss the virtual scaffolds generated by a full undirected scaffold decomposition as provided by the scaffold network approach. For example, a scaffold substructure relationship might connect a one-ring scaffold with a three-ring scaffold, not generating the additional two-ring scaffold between them that might be interesting for the analysis of bioactivity.

In summary, the aim of this study was the comparison of scaffold-tree- and scaffold-network-generated scaffolds for the mining of statistically significantly active scaffolds in primary screening data. In other words, we tested whether bioactive scaffolds could be predicted using chemical rules or whether the better strategy would be to explore all possible scaffolds and allow biology to determine the most relevant scaffolds.

■ METHODS

Scaffold Classification Procedure. *ST Algorithm.* The original scaffold tree algorithm was developed to provide a hierarchical scaffold decomposition with an emphasis on identifying the chemically most interesting and characteristic scaffolds.[12] These scaffolds could be used for a treelike representation of the scaffold chemical space of a library of compounds. Examples of scaffold tree decompositions are given in Figure 1 for three setrons.

In brief, a scaffold tree is generated as follows: Starting from a molecule, the first step is to generate a Murcko scaffold by removing all terminal side chains. In contrast to the Murcko definition of a scaffold, exocyclic and exolinker double bonds are retained, because of the different geometries imposed by the double bonds compared to sp^3 substituents. Starting from the Murcko scaffold as a child scaffold, additional parent scaffolds are calculated by iterative removal of peripheral rings, until no further rings can be removed, and the resulting scaffold is defined as level 1 of the scaffold hierarchy (in most cases, a one-ring scaffold). The scaffold hierarchy level of each child scaffold is calculated from the parent level + 1. The selection of which scaffold to retain for each of the levels is made based on a complex set of rules given in ref 12, aimed at selecting scaffolds that are chemically characteristic and favored in terms of medicinal chemistry over those that are more ubiquitous or less interesting in terms of medicinal chemistry. More precisely, the rules were designed to prioritize scaffolds with few acyclic linker bonds, containing spiro or bridged ring systems, ring systems of uncommon ring size (not 3, 5, or 6), rings with many heteroatoms, or nonaromatic saturated rings.

As an illustration, we describe the decomposition of Ondasetron in Figure 1 in more detail (in Figure 1, all blue scaffolds were generated by the ST and SN approaches, whereas green scaffolds were generated by the SN approach): At hierarchy level 4, the scaffold decomposition starts with the Murcko scaffold that was generated directly from Ondasetron by removal of terminal side chains. Scaffold decomposition through the removal of terminal rings generates the two scaffolds shown at hierarchy level 3 (one blue and one green scaffold). Of these two scaffolds, the rule set selects the blue scaffolds for the ST approach based on the lower number of acyclic linker bonds found in the blue scaffold relative to the green scaffold. The blue level-3 scaffold then serves as the child scaffold for the next iteration of the ring pruning procedure for level 2. From level 2, the blue scaffold is selected based on the presence of one nonaromatic ring compared with the fully aromatic green scaffold. The blue level-2 scaffold can be dissected again into two one-ring scaffolds. At level 1, the blue pyrrole scaffold is selected over the green scaffold based on the higher number of ring heteroatoms (heteroatoms of exocyclic double bonds such as the carbonyl oxygen are not counted here).

The scaffold tree was calculated by an in-house tool based on the Molinspiration toolkit, as described in ref 12.

SN Algorithm. Scaffold networks were calculated based on a modified version of the publicly freely available scaffold tree generator software[17,18] that implements the original scaffold tree algorithm based on the CDK cheminformatics library.[25,26] The program was modified so that no rules were applied to select one particular scaffold for one molecule per hierarchy level. Instead, all branches were explored. This is also illustrated in Figure 1. For the three setrons in Figure 1, the SN approach generated the green scaffolds in addition to the blue scaffolds that were generated by the ST approach (and also by the SN approach).

Scaffold Calculation. To restrict the additional computational costs for exploring additional branches in the SN approach, an upper threshold of 10 rings for scaffolds to be processed was applied. Scaffolds with more than 10 rings were not processed at all, because we assumed that decomposition of such large and complex molecules might not lead to results comparable with smaller more druglike scaffolds. To make the scaffold sets from the SN and ST approaches more comparable, we also removed all scaffolds with more than 10 rings from the ST data set, as well as all other scaffolds that were not generated by the SN approach, in part coming from molecules with more than 10 rings or resulting from slightly different implementations of the ST and SN codes (in total, 119 and 246 scaffolds were removed by this procedure for PubChem data sets 893 and 1634, respectively). As the final step, all SMILES of SN scaffolds were converted from CDK SMILES to Molinspiration SMILES to enable a text-based comparison with the original implementation of the scaffold tree approach.

CSE. Compound set enrichment[22] was used to identify statistically significantly enriched scaffolds in primary screening data. CSE applies a nonparametric statistical hypothesis test (Kolmogorov—Smirnov test) together with a multiple-hypothesis test correction (Bonferroni correction) to assess the activity of scaffolds found within the compounds of a screening set. This facilitates the identification of series of active compounds directly from primary screening data without the need to set an arbitrary activity threshold that can distort the evaluation of series of compounds. Another point of interest is that the pooling of compounds into defined groups can compensate for the lack of replicates in a primary HTS assay. This reduces the error in the scaffold activity evaluation in proportion to the number of compounds per scaffolds

p-Value Computation. In the context of compound set enrichment, the Kolmogorov—Smirnov test is used to compare the activity distribution of compounds having a common scaffold to the activity distribution of all compounds tested in the primary screen. The test evaluates the hypothesis (the null hypothesis, H_0): "there is no difference in the activity distribution defined by compounds having scaffold S and the background distribution". To compute the probability associated with this hypothesis (p-value) by the KS test, the two activity distributions are transformed into empirical cumulative distribution functions (ECDFs). The maximum distance between these two curves corresponds to the KS statistic D_{max}, where D_{max} is used to compute the p-value associated with H_0 for a scaffold.

Table 1. Data Sets Used in the Experiments

AID[a]	target	column[b]	P(AC)[c]	NC[d]	NAC[e]	NS_ST[f]	NS_SN[g]
893	hydroxysteroid (17-beta) dehydrogenase 4	28	0.0764	73919	5650	16017	28909
1634	pyruvate kinase	22	0.0006	263679	154	50690	80937

[a] PubChem assay ID. [b] Field number of the PubChem bioassay that was used to simulate the primary screening. [c] Proportion of active compounds. [d] Number of compounds. [e] Number of active compounds (according to the PubChem annotation). [f] Number of scaffolds with at least two compounds according to the scaffold tree approach. [g] Number of scaffolds with at least two compounds according to the scaffold network approach.

Table 2. Number of Scaffolds According to the Scaffold Tree (ST) and Scaffold Network (SN) Level for Scaffolds with at Least Two Compounds

		levels										
AID[a]	algorithm	all	1	2	3	4	5	6	7	8	9	10
893	SN	28909	637	8313	12016	5849	1505	387	145	48	8	1
	ST	16017	531	3338	6860	4027	1025	162	49	18	6	1
	ST/SN	1.80	1.20	2.49	1.75	1.45	1.47	2.39	2.96	2.67	1.33	1.00
1634	SN	80937	1115	24445	35284	15972	3451	532	109	21	7	1
	ST	50690	908	9979	23198	13132	2977	391	81	16	7	1
	ST/SN	1.60	1.23	2.45	1.52	1.22	1.16	1.36	1.35	1.31	1.00	1.00

[a] PubChem assay ID.

This p-value for a one-sided KS test is computed as

$$p = D_{\max} \sum_{j=0}^{[n(1-D_{\max})]} \left[\binom{n}{j} \left(1 - D_{\max} - \frac{j}{n}\right)^{n-j} \left(D_{\max} + \frac{j}{n}\right)^{j-1} \right] \quad (1)$$

where $D_{\max}$ is the maximum difference between the scaffold and the background ECDFs, n is the number of compounds, and $[n(1 - D_{\max})]$ is the greatest integer contained in $n(1 - D_{\max})$.

The p-value is influenced by two parameters: $D_{\max}$ and the number of compounds in the evaluated set. $D_{\max}$ quantifies the difference in the activities between the compounds of a scaffold and the background distribution. A higher $D_{\max}$ value results in smaller p-values. In addition, the number of compounds in a set influences the confidence in the differences observed in the activities. Larger numbers of active compounds lead to lower p-values and vice versa. Thus, if two scaffolds have the same $D_{\max}$ value but a different number of compounds, the scaffold with the largest number of compounds will have a lowest p-value. For scaffolds generated by both the ST and SN approaches, the ST scaffold is often less populated with compounds compared to the SN scaffold, increasing the chance of identifying significant signals by the SN approach.

The smaller the p-value, the higher the probability that the scaffold has a different distribution of activity readouts than the complete screening set. If the p-value is smaller than a critical level of significance ($\alpha = 0.01$), the null hypothesis is rejected, and the scaffold is considered as having a different activity distribution than the background distribution. That is, the scaffold is considered to be "active". We used a one-sided KS test to differentiate agonistic from antagonistic effects. This approach is applicable for one individual scaffold at a time. As a large number of scaffolds are tested in a primary screen, a multiple-hypothesis test correction was applied.

Multiple-Hypothesis Test Correction. The p-value can also be interpreted in a different way: the p-value represents the probability that the statistical test rejects the null hypothesis by chance. This means that, for each test, the probability that a real inactive scaffold is evaluated as active is equal to 0.01 (false positive). As the number of scaffolds evaluated can be extremely large, the number of false positives can also be significant. To correct this effect, a multiple-hypothesis test correction called the Bonferroni correction was applied. This correction adapts the level of significance to the number of scaffolds evaluated (level of significance divided by number of scaffolds). The consequence is that the probability of obtaining a false positive for all scaffolds tested, as opposed to each scaffold separately, is equal to 0.01. This correction requires the scaffolds to be independent, which is why the Bonferroni correction is applied separately to each hierarchy level in the ST and SN approaches. For the SN approach, different scaffolds at the same level can be populated by common compounds. These scaffolds are not independent, and in this case, the Bonferroni correction might be pessimistic. We considered scaffolds that were still significantly active after the Bonferroni correction as the active scaffolds.

Data Sets. Two data sets with quantitative HTS results were selected for the analysis. An overview over the data sets is given in Table 1.

RESULTS AND DISCUSSION

Comparison of Scaffolds Generated by the ST and SN Approaches. Additional significantly active scaffolds can be identified based on two mechanisms using scaffold networks: (1) Scaffolds not generated by the scaffold tree approach might be significantly active according to the KS statistics. (2) For a scaffold scarcely populated with compounds by the ST approach, additional active and/or weakly active molecules having the scaffold might be identified by the SN approach, reaching a critical number to be considered significant. Before considering the possibly more target- and assay-dependent comparison of the numbers of active scaffolds identified by each of the scaffold generation algorithms, we were first interested in comparing the ST and SN approaches in terms of the two major parameters underlying the

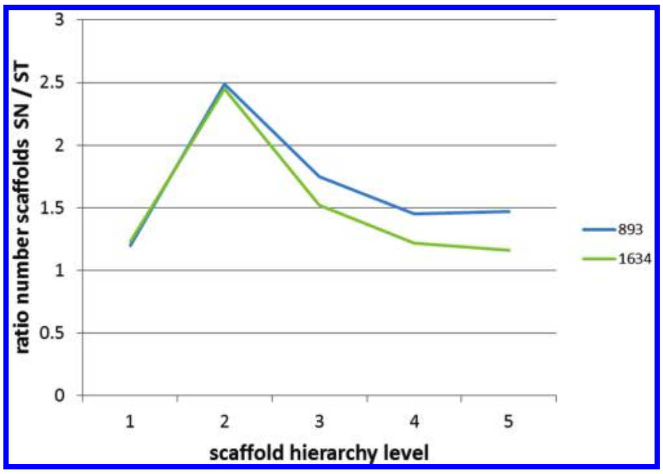


Figure 2. Relative numbers of scaffolds generated by the scaffold network (SN) approach compared to the scaffold tree (ST) approach for each scaffold hierarchy level for assays 893 and 1634.

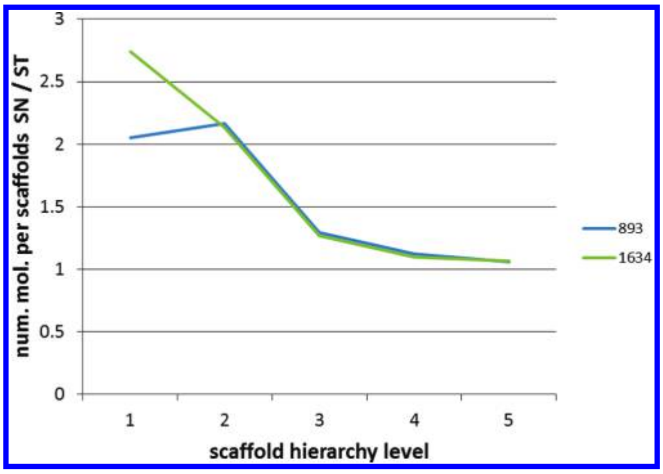


Figure 3. Relative numbers of molecules linked to the scaffolds generated by the SN approach compared to those generated by the ST approach for each scaffold hierarchy level for assays 893 and 1634.

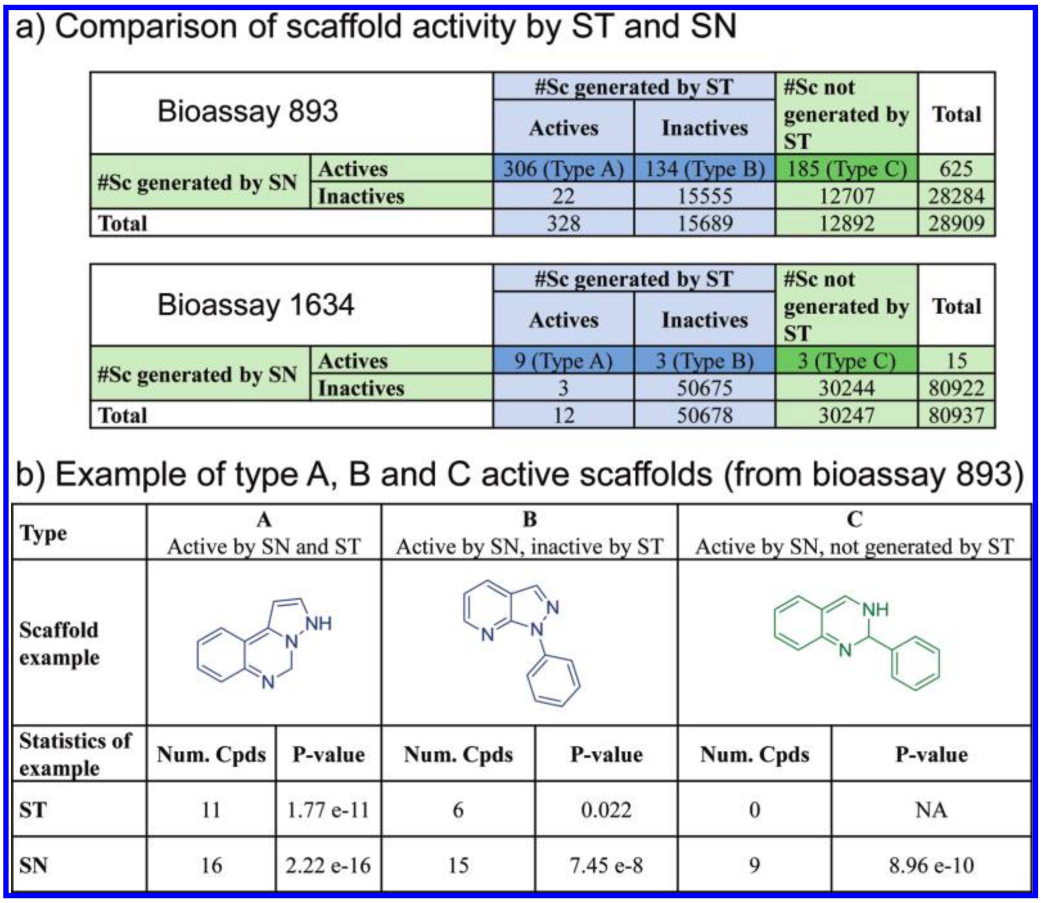
two mechanisms of identifying additional active scaffolds: the numbers of scaffolds generated by the ST and SN approaches and the numbers of molecules represented by each of the scaffolds generated by the ST or SN approach. The former parameter determines the number of unique scaffolds examined for activity, and the latter parameter determines the number of molecules and measured data points considered for the statistical analysis of a scaffold.

As a first step, scaffolds were generated for the two PubChem data sets using the scaffold tree and scaffold network algorithms. The numbers of scaffolds generated, also counted separately per scaffold tree level, are listed in Table 2. Interestingly, the ratio of the numbers of scaffolds generated by the SN approach to the number generated by the ST approach was found to be very similar between the two data sets (see Figure 2). Averaged over all hierarchy levels, this ratio was found to be 1.7 ± 0.2. Accordingly, a large number of additional scaffolds were generated by the scaffold network approach, compared to the scaffold tree algorithm. Considering the different hierarchy levels, a clear maximum of this ratio could be found for hierarchy level 2 (ca. 2.5 times more scaffolds by the SN versus the ST approach). For hierarchy level 1, only a few additional scaffolds were generated ($SN/ST \approx 1.2$), and from level 3 on, the SN/ST ratio decreased again for both data sets until hierarchy level 5. Starting from level 6, the SN/ST ratios behaved differently for the two PubChem data sets. This is likely because of the low number of scaffolds found at these levels. This analysis suggests that the SN algorithm generates a large number of additional scaffolds, in particular for hierarchy levels from 2 to 5. Molecules of this size are also most interesting for the development of novel druglike molecules.

As the next step, we analyzed the numbers of molecules that were associated with each scaffold, that is, the number of molecules from which a particular scaffold could be derived by the ST or SN approach. The SN algorithm explores all possible ring-pruning paths from the starting molecules to generate smaller scaffolds, whereas the ST algorithm explores only one such path using the ring-pruning rules. Therefore, on average, more molecules should be associated with each scaffold resulting from SN- versus ST-generated scaffolds. Here, we were interested in how many additional molecules were associated with a scaffold using the SN compared to the ST approach. This was expressed using the ratio of molecules associated with a scaffold using the ST versus the SN approach. This ratio of the number of molecules per scaffold was calculated individually for each scaffold and averaged over all scaffolds of the data set that were generated by both the ST and SN approaches. As with the ratios of the numbers of scaffolds, the ratios of the numbers of molecules per scaffolds between the SN and ST approaches were found to be very similar for the two PubChem data sets (see Figure 3), with values of 1.44 ± 1.91 and 1.41 ± 3.64 for data sets 893 and 1634, respectively. In contrast to the numbers of scaffolds, the largest numbers of additional molecules per scaffolds (with ratios of >2) were found for hierarchy levels 1 and 2. For level 3, we found a significant decrease in additional molecules per scaffold for all data sets to about 1.28, and starting from hierarchy level 4, only a few additional molecules were identified. This observation is not surprising: Because the scaffolds get much more specific at higher hierarchy levels, it is less likely that such scaffolds that were also generated by the ST approach might find additional samples starting from additional molecules.

Averaged over all levels, these findings demonstrate that a large number of additional scaffolds were generated by the SN approach compared to the ST approach. In addition, the scaffolds common to the two algorithms represent a significantly larger number of molecules using the SN approach compared to ST and, hence, also have more associated biological activity data per scaffold. Therefore, using the SN approach might lead to statistically more significant results, which might also lead to the identification of even weaker signals than using the ST approach.

Mining for Active Scaffolds. We demonstrated in our previous publication[22] that compound set enrichment using the scaffold tree algorithm to generate scaffolds is able to identify most of the scaffolds that are classified as active according to a binomial hypothesis test based on the PubChem compound annotation (active, inconclusive, or inactive, where we also considered inconclusive as inactive for this study). CSE identified not only most active scaffolds but also a large number of scaffolds that were not identified by the PubChem-annotation-based test. Novel active scaffolds according to CSE often contained many inconclu sive compounds that showed weak but clearly visible effects on the concentration−response curves and were therefore considered as true active scaffolds, populated with weakly active compounds.



NH
N
N
N
N
N
NH
N

a) Comparison of scaffold activity by ST and SN

Bioassay 893		#Sc generated by ST		#Sc not generated by ST	Total
		Actives	Inactives		
#Sc generated by SN	Actives	306 (Type A)	134 (Type B)	185 (Type C)	625
	Inactives	22	15555	12707	28284
Total		328	15689	12892	28909

Bioassay 1634		#Sc generated by ST		#Sc not generated by ST	Total
		Actives	Inactives		
#Sc generated by SN	Actives	9 (Type A)	3 (Type B)	3 (Type C)	15
	Inactives	3	50675	30244	80922
Total		12	50678	30247	80937

b) Example of type A, B and C active scaffolds (from bioassay 893)

Type	A Active by SN and ST		B Active by SN, inactive by ST		C Active by SN, not generated by ST	
Scaffold example						
Statistics of example	Num. Cpds	P-value	Num. Cpds	P-value	Num. Cpds	P-value
ST	11	1.77 e-11	6	0.022	0	NA
SN	16	2.22 e-16	15	7.45 e-8	9	8.96 e-10

Figure 4. Comparison of scaffold activity as determined by the ST and SN approaches. (a) Detailed comparisons of the numbers of active scaffolds (#Sc) identified by the ST and SN approaches for PubChem bioassays 893 and 1634. In both bioassays, significant numbers of additional active scaffolds were identified by the SN approach compared to the ST approach (for assay 893, 625 by SN versus 328 by ST; for assay 1634, 15 by SN versus 12 by ST). Active SN scaffolds were identified from three different categories, highlighted in the tables: type A (active by SN and ST), type B (active by SN, inactive by ST), and type C (active by SN, not generated by ST). Almost all scaffolds active by the ST approach were also active by the SN approach (type A). (b) Examples of type-A, -B, and -C active scaffolds (from bioassay 893). The type-A example scaffold has 11 and 16 associated compounds from the ST and SN approaches, respectively, leading to significant *p*-values for both the ST and SN approaches. The type-B example has fewer molecules associated with the scaffold by the ST approach (6) and still 15 identified by the SN approach, leading to a significant *p*-value for the SN but not the ST approach. The example scaffold of type C was not generated by the ST approach (0 molecules associated with the scaffold), so it could be active only by the SN approach.

Here, we were interested in the comparison of the classification of ST- versus SN-generated scaffolds, in particular if and how additional active scaffolds might be identified by the scaffold network approach over the scaffold tree approach. To answer this question, compound set enrichment calculations were performed using ST- and SN-generated scaffolds based on the results from PubChem assays 893 and 1634. The results for the two PubChem data sets are shown in Figure 4.

The CSE results shown in Figure 4a are comparable between the two PubChem data sets. In both cases, significantly more active scaffolds were identified by the SN approach compared to the ST approach (for assay 893, 625 by SN versus 328 by ST; for assay 1634, 15 by SN versus 12 by ST). As a consequence, only a fraction of the scaffolds identified as active by the SN approach were also identified by the ST approach (for assay 893, 306 by ST/625 by SN = 0.49; for assay 1634, 9 by ST/15 by SN = 0.6), meaning that about 50% and 40% of the active scaffolds from the SN approach would be missed if ST scaffolds were analyzed alone. Such scaffolds active by both the SN and ST approaches are denoted type-A scaffolds. As mentioned earlier, additional significantly active scaffolds could be found by two mechanisms: (1) For a scaffold scarcely populated with weakly active scaffolds from the ST approach, additional weakly active molecules having the scaffold might be identified by the SN approach, reaching a critical number to be considered significant (type B). (2) Scaffolds not generated by the scaffold tree approach might be significantly active according to the KS statistics (type C). Figure 4b shows examples for each of the three types of active SN scaffolds. The PubChem data sets show that both mechanisms identified additional active scaffolds. No clear preference between the two mechanisms was observed, however. Whereas for assay 893 more additional active scaffolds were identified among the scaffolds that were generated only by the SN algorithm, more additional active scaffolds were found for assay 1634 among the scaffolds that were considered inactive according to the *p*-values of the ST scaffolds. Interestingly, for both data sets, there were also small numbers of active scaffolds that were considered active only by the ST approach and not by the SN approach. For these examples, it seems that having more bioactivity data associated with a scaffold for the statistical analysis, resulting from the scaffold network

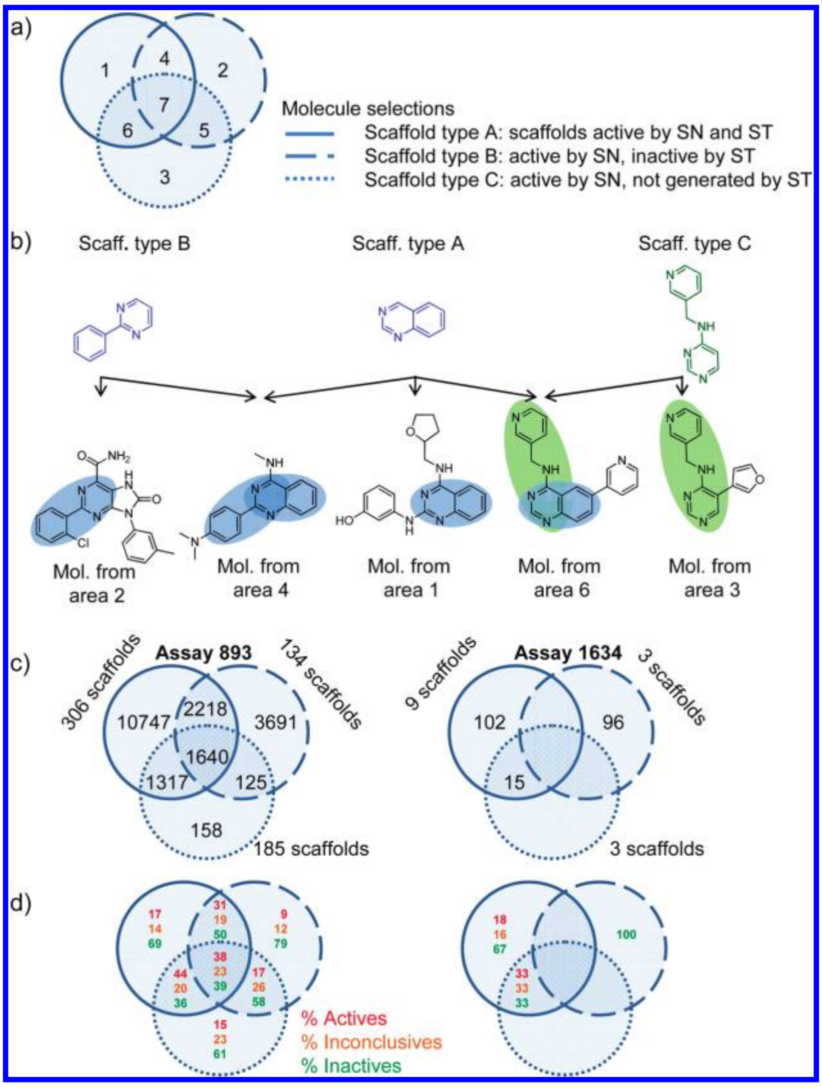


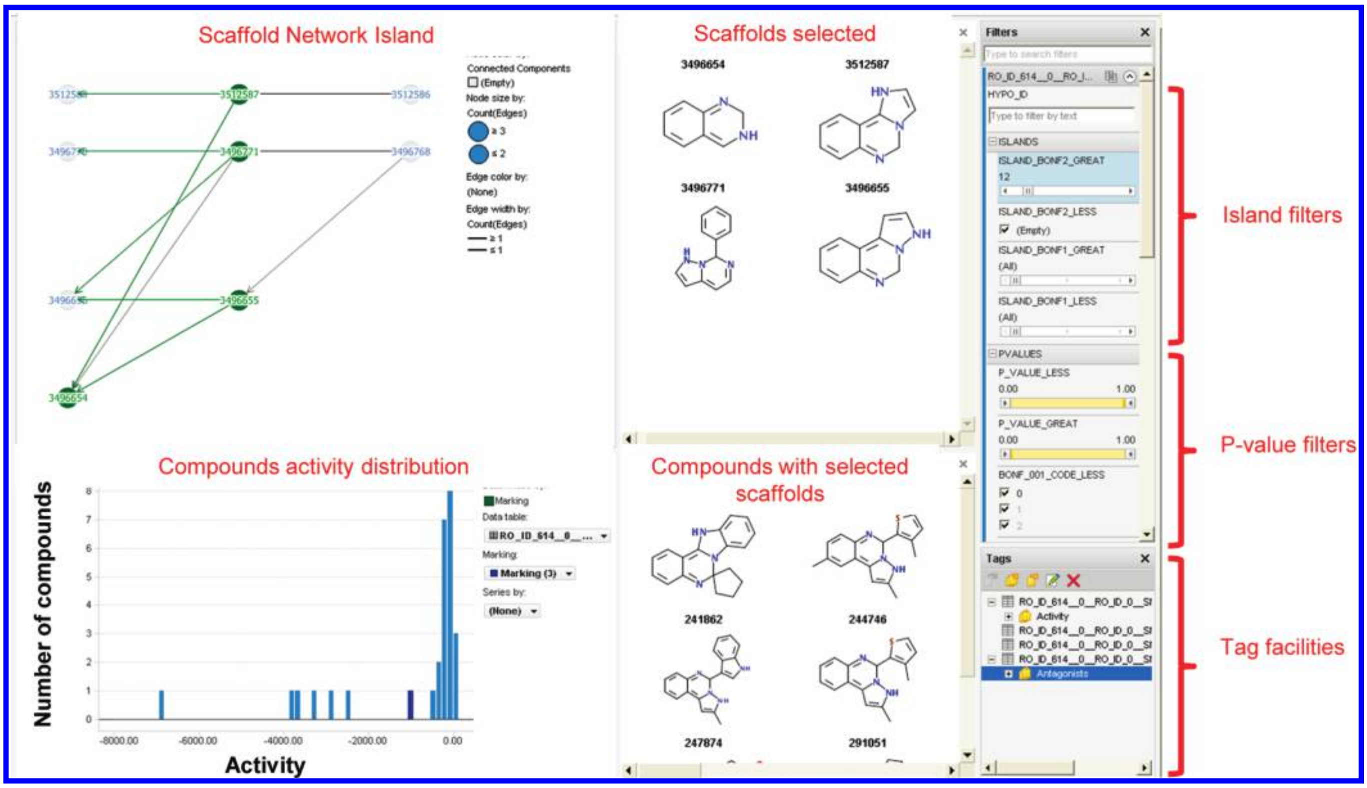
Figure 5. Overlap analysis of molecule selections from active scaffolds of types A, B, and C. (a) Overview and definition of the overlapping areas in a Venn diagram representation of the three scaffold types. Molecules in areas 2, 3, and 5 would be identified only by the SN approach through types B and C active scaffolds and would be missed by the ST approach. (b) Examples of active molecules from bioassay 893 containing active scaffolds of only one type (example molecules from areas 1−3) and molecules containing scaffolds of different types (here, the example molecule from area 4 contains two active scaffolds of types A and B, and the example molecule from area 6 contains two active scaffolds of types A and C). The example molecules from areas 2 and 3 are selected only by SN scaffolds (of types B and C), whereas the example molecules from areas 4 and 6 also contain a type-A scaffold in addition to a type-B or -C scaffold. Therefore, the latter molecules would also be selected by the ST approach, even though they contain active scaffolds identified only by the SN approach. (c) Numbers of molecules of assays 893 and 1634 selected by scaffolds of types A−C in the different areas of the Venn diagram. (d) Percentages of actives, inconclusives, and inactives according to the PubChem annotation based on concentration response curves found in the different areas of the Venn diagram.

algorithm, caused more inactive molecules to be added, resulting in inactive scaffolds.

In summary, compound set enrichment by SN scaffolds identified a significant number of novel active scaffolds compared to the ST approach, by both type-B (providing more samples for existing ST scaffolds) and type-C (novel scaffolds not generated by the ST approach) mechanisms.

Mining for Active Molecules. For most practical applications, one would reach a stage of the analysis where actual molecules containing a scaffold would have to be selected and further characterized. The major question for selecting molecules is whether the additional active scaffolds identified by the SN approach are linked to additional active molecules that would not be identified using ST scaffolds alone or whether the SN scaffolds mainly identify the same molecules as the ST approach? With the SN algorithm, in contrast to the ST algorithm, each molecule has several scaffolds on each scaffold tree level. This has the advantage that greater amounts of bioactivity data are available for each scaffold for the statistical analysis. On the other hand, each molecule might also be selected by multiple active scaffolds on each scaffold tree level. In the worst case, this might result in a scenario in which the SN approach identifies a large set of additional active scaffolds but no additional active molecules, as all active molecules were already linked to active scaffolds from the ST approach.

To demonstrate the usefulness of the scaffold network approach not only for the selection of active scaffolds but also for the selection of active molecules, we analyzed the overlap of the molecules selected by active scaffolds of types A, B, and C.



Filters

RO_ID_614__0__RO_I...

HYPO_ID

ISLANDS

ISLAND_BONF2_GREAT 12

ISLAND_BONF2_LESS ☑ (Empty)

ISLAND_BONF1_GREAT (All)

ISLAND_BONF1_LESS (All)

PVALUES

P_VALUE_LESS 0.00 1.00

P_VALUE_GREAT 0.00 1.00

BONF_001_CODE_LESS ☑ 0 ☑ 1 ☑ 2

Tags

RO_ID_614__0__RO_ID_0__SI

Activity

RO_ID_614__0__RO_ID_0__SI

RO_ID_614__0__RO_ID_0__SI

RO_ID_614__0__RO_ID_0__SI

Antagonists

Figure 6. Visualization of a scaffold network island of significantly active scaffolds from assay 893 using a Spotfire session. Scaffolds are arranged based on the scaffold hierarchy. Islands can be switched using the filter panel. Clicking on nodes or edges in the network window displays the structures of corresponding scaffolds and compounds. The activity distribution of compounds is also shown.

Figure 5a gives a definition of the different possible combinations of overlapping scaffolds, Figure 5b provides an example for molecules matching scaffolds from only one or different scaffold types. For this analysis, we considered only scaffolds from hierarchy level 2 or higher with the assumption that single-ring scaffolds, the majority of scaffolds in level 1, might be too unspecific for the general evaluation intended here. However, in any particular case, such scaffolds might still be of value. Removal of level-1 scaffolds removed 58 of 647 active scaffolds for assay 893 and 4 of 14 scaffolds for assay 1634. An overview of the different selections of molecules by the various scaffolds is given in Figure 5c and 5d.

For both assays, the largest numbers of selected molecules came from type-A scaffolds (Figure 5c). The second largest set of molecules was selected by the type-B scaffolds. For both of these sets, large fractions of molecules were selected that were found only by the scaffolds of types A and B, respectively (areas 1 and 2 of the Venn diagram, Figure 5c). In contrast, for type-C scaffolds (SN active, ST not generated), only relatively few molecules were selected of which most were also selected by scaffolds from other groups (e.g., areas 5 and 6 for assay 893 and area 6 for assay 1634, Figure 5c). Interestingly this lower number of molecules selected by type-C scaffolds is not reflected in a lower number of active type-C scaffolds, compared to type-A and -B scaffolds; for example, for assay 893, there are even more type-C scaffolds (185) than type-B scaffolds (134). A likely explanation for this finding might be that scaffolds that are found in only a few molecules have a higher chance of not being generated by the ST algorithm than scaffolds that are found in many diverse molecules.

Novel molecules that were selected only by the SN approach and not by the ST approach are found in areas 2, 3, and 5 of the Venn diagram visualization (Figure 5c). For both assays, a significant number of additional novel molecules were identified by the SN approach. For assay 893, in addition to the 15922 compounds also identified by the ST approach, 3974 (25%) compounds were identified by the SN approach alone from areas 2, 3, and 5. For assay 1634, 96 novel compounds were identified in addition to the 117 from the ST approach alone, increasing the number of compounds by 82%. As mentioned above, the majority of these additional compounds were identified using type-B scaffolds and only to a lower degree using type-C scaffolds.

The last important point to be addressed in the overlap analysis was whether the novel identified compounds were also active in terms of showing dose−response effects. Therefore, we identified the selected molecules in each of the areas of the Venn diagram by their PubChem annotation (which was based on dose−response data) as active, inconclusive, or inactive and plotted the fraction of each activity annotation in the areas of the Venn diagram (Figure 5d). Significant numbers of actives were found in all areas of the Venn diagram containing selected molecules (except assay 1634, area 2). A weak trend in the data might indicate that there are larger fractions of actives in overlapping areas 4−7 compared to the areas unique for one scaffold type. This would make sense, because molecules from overlap areas in the Venn diagram contain more than one active scaffold and might therefore also be more likely to be active. Another interesting point is that there are also significant numbers of inconclusive compounds selected in most areas of the Venn diagram. In our original CSE publication, we showed that most of the selected inconclusive compounds exhibited real but weak activities.[22] In practice, the statistical analysis might

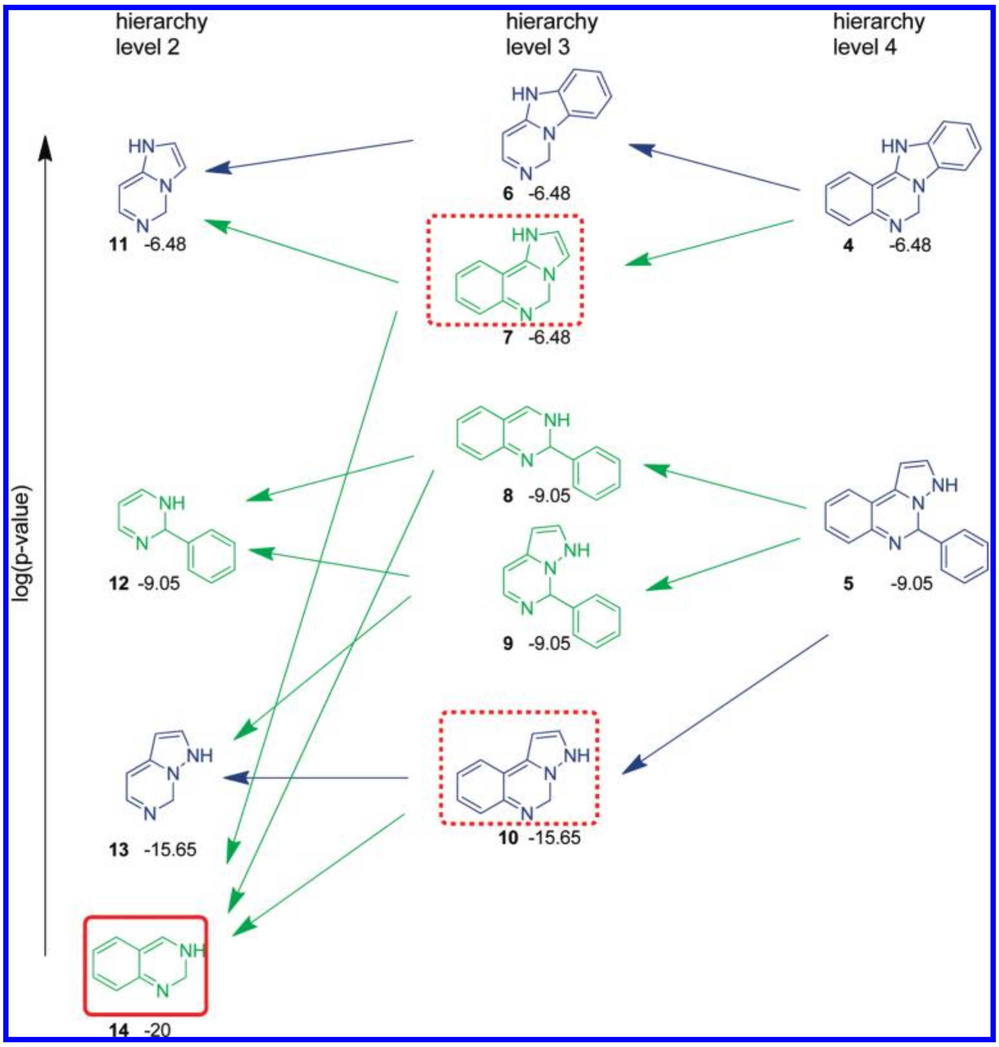


Figure 7. Scaffold network island of significantly active scaffolds from assay 893. Scaffolds are arranged based on the scaffold hierarchy level (x axis) and approximately by log(p-value) (y axis). Blue scaffolds were generated by both the ST and SN approaches, and green scaffolds were generated only by the SN approach. Scaffolds from hierarchy level 1 were omitted for scaffold islands, because they often led to meaningless connections. Scaffold **14** in the red box is the link between the upper and lower parts of the network. The two scaffolds in dashed boxes are the direct child scaffolds of **14** with best p-values from each sub-island, All shown scaffolds are statistically active according to the SN approach. All blue scaffolds except **5** are also statistically active according to the ST approach.

enable the separation of weak inconclusives from real false positives.

In summary, our data indicates that the selection of compounds based on active scaffolds identified by CSE benefits from the use of the SN approach compared to the ST approach. A significant number of additional molecules were identified that would not have been identified by the ST approach alone.

Visualization of Scaffold Network Islands. For the visual analysis of screening data sets, we have developed a network representation of the scaffold network approach. For large data sets, however, visualizing the entire network with all nodes (scaffolds) and all edges is much too large and complex to be helpful in the analysis of trends and relationships in the activity of scaffolds. Therefore, we developed the concept of scaffold network islands: subnetworks consisting only of active scaffolds connected by parent—child relationships. In this way, smaller islands of active and related scaffolds can be analyzed one after the other, possibly allowing extraction of initial structure—activity relationship (SAR) information. Hierarchy-level-1 scaffolds were not considered for this analysis because they often led to meaningless connections of groups of scaffolds, for example, with highly ubiquitous rings such as pyridine or furan. The aggregation of active scaffolds into islands starts by selecting all active scaffolds at level 2 as island starting points. Then, all active scaffolds at level 3 that are connected to active scaffolds at level 2 are added to the parent's island. If a child scaffold has more than one active parent scaffold, the two corresponding islands are merged. If a child scaffold is considered inactive after the Bonferroni correction but has a smaller p-value than its active parent, it is still added to the parent's island. The latter procedure corrects for an artifact of the fact that the Bonferroni correction is applied separately for each hierarchy level and is therefore more stringent for hierarchy levels with large number of scaffolds. Finally, all active scaffolds at level 3 that are not yet part of an island are considered as first nodes of new islands. This procedure is reapplied iteratively until the highest available hierarchy level (here, level 10) is reached. After all active islands have been identified, they are sorted by the number of active scaffolds in the island (a global p-value for the full island could also be computed, but it would be strongly influenced by scaffolds at the lowest levels). In this study, we considered only active scaffolds that were still active after the Bonferroni correction as starting points for islands. For assays with very few active molecules or very moderately active molecules, it might be interesting to consider scaffolds with small *p*-values that are inactive after the Bonferroni correction. In this case, we suggest that one consider as a starting point all scaffolds having a *p*-value smaller than or equal to some *p*-value threshold (e.g., 0.01).

We used Spotfire[27] to visualize the scaffold network and the islands of active scaffolds. Figure 6 shows a screenshot of an example session of a scaffold network island of active scaffolds from assay 893. Each node in the network represents a scaffold, and each edge in the network represents a parent−child relationship. The scaffold nodes are arranged by their hierarchy level on the *x* axis (with lower to higher levels from left to right) and by their *p*-values on the *y* axis [$\log_{10}$(*p*-value), with larger, less significant values to smaller, more significant values from top to bottom). This interface is interactive, and by clicking on nodes or edges, one can display corresponding scaffold structures, compounds with these scaffolds, and the compound's activity distribution in separate windows. A panel filter is available on the right of the session to browse from island to island (or to filter results by *p*-value, compound activity, etc.). Finally, a tag panel enables scaffold annotations during the analysis. The highly flexible Spotfire platform allows one to easily generate specifically adapted and enhanced versions of the shown example.

Figure 7 gives a more detailed overview of the island of active scaffolds from assay 893 that is introduced in Figure 6. This scaffold network island consists of 11 active scaffolds according to the SN approach. In this visualization, it is immediately apparent that 2,3-dihydroquinazoline scaffold **14** (in the red box) is the most significant active scaffold. Scaffold **14** is the only link connecting the otherwise unconnected subislands derived from scaffolds **4** and **5**, and therefore, it represents the largest number of active molecules within the scaffolds of the island. From the statistical analysis, it is also apparent that 2,3-dihydroquinazoline is also specific enough not to represent too many inactive molecules.

Notably, scaffold **14** was generated only by the SN approach and would be missed if only the ST algorithm were used. In addition, as mentioned above, **14** is the only node connecting two subnetworks of the network island, generated by scaffolds derived from **4** and **5**. This link between the subislands would also be missed by the ST approach alone. Of the two direct child scaffolds **7** and **10** (red dashed boxes), scaffold **7** was also generated only by the SN approach.

When looking for other scaffold SAR, classifying scaffolds by a similarity-based approach such as clustering might be useful. This scaffold clustering could be applied separately to each level or for scaffolds with a common parent. The advantage is to compare the activity or to aggregate similar scaffolds together. However, as for the SN approach, each scaffold can have several parents, a scaffold could be classified in different clusters, and the results might be confusing. However, if a parent scaffold of interest has been identified, clustering its scaffold children on the fly might be very useful.

CONCLUSIONS

In this work, we have introduced the scaffold network approach as an extension of the scaffold tree approach to explore the full scaffold space that is defined by the molecules of a library, instead of a rule-based subselection of this space as with the scaffold tree approach. Whereas the scaffold tree approach with the rule-based selection of scaffolds has a clear focus on identifying chemically interesting scaffolds within a library, for example, as applied for the prioritization of commercial vendor compounds, the focus of the scaffold network approach is on the analysis of biological data linked to chemical space. Therefore, the exploration of the full scaffold space is important for identifying any scaffold that shows a desired biological effect irrespective of the level of interest in the scaffold's chemistry.

We compared the performance of the scaffold tree and scaffold network approaches for the identification of statistically significantly active scaffolds from primary screening data. Using the scaffold network approach, about twice as many active scaffolds could be identified as with the scaffold tree approach. The additional scaffolds were identified by two mechanisms: First, additional sampling of a scaffold's bioactivity was achieved by linking a scaffold to each molecule comprising this scaffold as a substructure, instead of linking it to only a subselection of these molecules selected by the scaffold tree algorithm (referred to as type-B scaffold in the Results and Discussion section). Second, additional active scaffolds were found by exploration of additional scaffolds that were not generated by the scaffold tree approach (referred to as type-C scaffolds in the Results and Discussion section). The two mechanisms led to comparable numbers of additional active scaffolds, whereas the former mechanism intrinsically led to scaffolds having more active molecules compared to the scaffolds selected by the latter mechanism. If active scaffolds were used to select molecules containing these scaffolds, the novel scaffolds identified by the SN approach were able to select a large number of additional active molecules that would not have been selected by active ST scaffolds alone.

We propose visualizing scaffold networks using the concept of smaller locally connected islands of active scaffolds. This provides an overview of the SAR of the statistical significance of the activity observed with similar scaffolds related to each other by parent−child relationships.

The network graphs of full scaffold networks are much more complex and more challenging to lay out than the tree graphs of the scaffold tree. When some activity data are present to select edges of interest in the scaffold network graph and thus decompose the scaffold networks into activity islands, one circumvents the problem of a full layout of the complete network graph. However, in place of activity data, the scaffold tree approach provides heuristic selection criteria for edges of interest based on the chemical structure within the complete scaffold network and thus obtains a simplified network graph more amenable to layout and browsing than the full network.

In conclusion, we recommend the scaffold network approach for the analysis of bioactivity linked to chemical space and the scaffold tree approach for the analysis of the chemistry within a library of molecules.

AUTHOR INFORMATION

Corresponding Author

*E-mail: steffen.renner@novartis.com. Tel.: +41 61 32 48879. Fax: +41 61 32 46261.

ACKNOWLEDGMENT

We gratefully thank the NIBR Education Office for providing the fellowship of T.V. We thank Edgar Jacoby, Christian Parker, Hanspeter Gubler, Ji-Hu Zhang and the MLI group for helpful discussions.

REFERENCES

(1) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–93.

(2) Krier, M.; Bret, G.; Rognan, D. Assessing the scaffold diversity of screening libraries. *J. Chem. Inf. Model.* **2006**, *46*, 512–524.

(3) Medina-Franco, J.; Martínez-Mayorga, K.; Bender, A.; Scior, T. Scaffold Diversity Analysis of Compound Data Sets Using an Entropy-Based Measure. *QSAR Comb. Sci.* **2009**, *28*, 1551–1560.

(4) Lee, M. L.; Schneider, G. Scaffold architecture and pharmacophoric properties of natural products and trade drugs: Application in the design of natural product-based combinatorial libraries. *J. Comb. Chem.* **2001**, *3*, 284–289.

(5) Grabowski, K.; Baringhaus, K.-H.; Schneider, G. Scaffold diversity of natural products: Inspiration for combinatorial library design. *Nat. Prod. Rep.* **2008**, *25*, 892–904.

(6) Schneider, G.; Schneider, P.; Renner, S. Scaffold-Hopping: How Far Can You Jump? *QSAR Comb. Sci.* **2006**, *25*, 1162–1171.

(7) Brown, N.; Jacoby, E. On scaffolds and hopping in medicinal chemistry. *Mini Rev. Med. Chem.* **2006**, *6*, 1217–1229.

(8) Hu, Y.; Bajorath, J. Structural and Potency Relationships between Scaffolds of Compounds Active against Human Targets. *ChemMedChem* **2010**, *5*, 1681–1685.

(9) Hu, Y.; Bajorath, J. Scaffold distributions in bioactive molecules, clinical trials compounds, and drugs. *ChemMedChem* **2010**, *5*, 187–190.

(10) Wilkens, S. J.; Janes, J.; Su, A. I. HierS: Hierarchical scaffold clustering using topological chemical graphs. *J. Med. Chem.* **2005**, *48*, 3182–3193.

(11) Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. Charting biologically relevant chemical space: A structural classification of natural products (SCONP). *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17272–17277.

(12) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The scaffold tree—Visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inf. Model.* **2007**, *47*, 47–58.

(13) Ertl, P.; Schuffenhauer, A.; Renner, S. In *Chemoinformatics and Computational Chemical Biology*; Bajorath, J., Ed.; Humana Press: New York, 2010; pp 245−260.

(14) Molinspiration Molecule Clustering. http://www.molinspiration.com/docu/clusterer/ (accessed Feb 1, 2011)

(15) *MOE, Molecular Operating Environment*, version 2010.10; Chemical Computing Group Inc.: Montreal, Canada, 2010.

(16) Agrafiotis, D. K.; Wiener, J. J. M. Scaffold Explorer: An Interactive Tool for Organizing and Mining Structure−Activity Data Spanning Multiple Chemotypes. *J. Med. Chem.* **2010**, *53*, 5002–5011.

(17) ScaffoldHunter. http://scaffoldhunter.sourceforge.net (accessed Jan 23, 2011)

(18) Wetzel, S.; Klein, K.; Renner, S.; Rauh, D.; Oprea, T. I.; Mutzel, P.; Waldmann, H. Interactive exploration of chemical space with Scaffold Hunter. *Nat. Chem. Biol.* **2009**, *5*, 581–583.

(19) Renner, S.; van Otterlo, W. A, L.; Dominguez Seoane, M.; Möcklinghoff, S.; Hofmann, B.; Wetzel, S.; Schuffenhauer, A.; Ertl, P.; Oprea, T. I.; Steinhilber, D.; Brunsveld, L.; Rauh, D.; Waldmann, H. Bioactivity-guided mapping and navigation of chemical space. *Nat. Chem. Biol.* **2009**, *5*, 585–592.

(20) Wetzel, S.; Wilk, W.; Chammaa, S.; Sperl, B.; Roth, A. G.; Yektaoglu, A.; Renner, S.; Berg, T.; Arenz, C.; Giannis, A.; Oprea, T. I.; Rauh, D.; Kaiser, M.; Waldmann, H. A scaffold-tree-merging strategy for prospective bioactivity annotation of gamma-pyrones. *Angew. Chem., Int. Ed.* **2010**, *49*, 3666–3670.

(21) Antonchick, A. P.; Gerding-Reimers, C.; Catarinella, M.; Schürmann, M.; Preut, H.; Ziegler, S.; Rauh, D.; Waldmann, H. Highly enantioselective synthesis and cellular evaluation of spirooxindoles inspired by natural products. *Nat. Chem.* **2010**, *2*, 735–740.

(22) Varin, T.; Gubler, H.; Parker, C. N.; Zhang, J.-H.; Raman, P.; Ertl, P.; Schuffenhauer, A. Compound Set Enrichment: A Novel Approach to Analysis of Primary HTS Data. *J. Chem. Inf. Model.* **2010**, *50*, 2067–2078.

(23) Thompson, A. J.; Lummis, S.C. R. 5-HT$_3$ Receptors. *Curr. Pharm. Des.* **2006**, *12*, 3615–3630.

(24) Hu, Y.; Bajorath, J. Combining Horizontal and Vertical Substructure Relationships in Scaffold Hierarchies for Activity Prediction. *J. Chem. Inf. Model.* **2011**, *51*, 248–257.

(25) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493–500.

(26) Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. L. Recent developments of the chemistry development kit (CDK)—An open-source java library for chemo- and bioinformatics. *Curr. Pharm. Des.* **2006**, *12*, 2111–2120.

(27) Spotfire. http://spotfire.tibco.com/ (accessed Feb 14, 2011).