JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

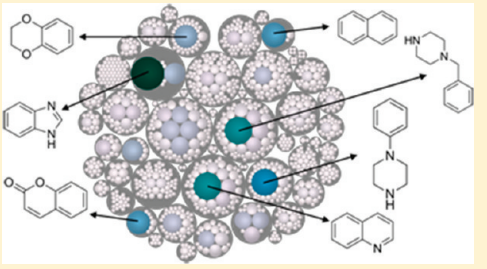
# Scaffold Diversity of Exemplified Medicinal Chemistry Space

Sarah R. Langdon,[†] Nathan Brown,*[,†] and Julian Blagg*[,†]

[†]Cancer Research UK Cancer Therapeutics Unit, The Institute of Cancer Research, 15 Cotswold Road, Sutton, Surrey SM2 5NG, U.K.

Supporting Information

**ABSTRACT:** The scaffold diversity of 7 representative commercial and proprietary compound libraries is explored for the first time using both Murcko frameworks and Scaffold Trees. We show that Level 1 of the Scaffold Tree is useful for the characterization of scaffold diversity in compound libraries and offers advantages over the use of Murcko frameworks. This analysis also demonstrates that the majority of compounds in the libraries we analyzed contain only a small number of well represented scaffolds and that a high percentage of singleton scaffolds represent the remaining compounds. We use Tree Maps to clearly visualize the scaffold space of representative compound libraries, for example, to display highly populated scaffolds and clusters of structurally similar scaffolds. This study further highlights the need for diversification of compound libraries used in hit discovery by focusing library enrichment on the synthesis of compounds with novel or underrepresented scaffolds.



## INTRODUCTION

Scaffold diversity is one of many parameters that may be used to characterize compound screening libraries.[1] The balance between the diversity of scaffolds within a library and the density of coverage for each scaffold varies according to the library design principles applied. Dense representation over small numbers of scaffolds is often applicable in libraries focused on a particular biological target class where thorough coverage of pharmacophore space is desired, for example in kinase-focused libraries.[2] However, such dense coverage of scaffold space may impart significant redundancy due to over population with structurally similar compounds. However, sparse representation of a large number of scaffolds may also be problematic in a screening library; for example, hit confirmation and rapid generation of structure activity relationships is challenging for compounds that are single exemplars of a particular scaffold. Thus the balance between scaffold diversity and scaffold representation is an important feature in library design and use.

In order to analyze the scaffold diversity of a compound library, a suitable representation of a scaffold is required. The definition of a scaffold often depends on the problem and the expertise of the individual defining the scaffold. One frequently applied description of a scaffold is the Markush structure, which first appeared in a patent, filed by Eugene A. Markush of the Pharma-Chemical Corporation in 1924.[3] The patent claimed a family of pyrazolone dyes and described a scaffold structure appended with "R" groups to denote the substitution patterns (Figure 1). Markush structures are generic and use variables to encode more than one structure in a single representation.

Markush structures are often used in patent applications to define the scope of a chemical series.[4] However, Markush structures often differ from how a medicinal chemist would define the relevant scaffold of a chemical series. A scaffold may, for example, define the core structure essential for pharmacological activity and the appended substituent vectors define optimal substitution patterns. For example, the HSP90 inhibitor NVP-AUY922 (Figure 2a)[5] is represented by a Markush structure (Figure 2b) in the corresponding patent application.[6] A medicinal chemistry representation of the scaffold may be more granular (Figure 2c) to reflect the importance of the resorcinol and isoxazole amide functionalities for pharmacological activity as well as the benzylic amine substituent for aqueous solubility.[5]

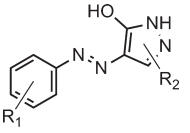



**Figure 1.** An interpretation of the Markush structure as described in the 1924 Markush patent.[3]

A preferred scaffold representation is objective, invariant, and is not data set dependent.[7] One such method is the Murcko framework, proposed by Bemis and Murcko in 1996 which has been used to analyze the structures of known drugs.[8] The method dissects molecules into ring systems (Figure 2d), linkers (Figure 2e), side chain atoms (Figure 2f), and the framework (Figure 2g), which is the union of ring systems and linkers in a molecule. A Murcko framework (Figure 2h) retains information on atom type, whereas a graph framework[8] (Figure 2i) reduces all atoms to carbon and all bonds to single bonds.

There are examples of methods where the scaffold definition is data set dependent, such as a Maximum Common Substructure (MCS) search. In this approach, molecules are typically clustered

**Received:** March 25, 2011
**Published:** August 31, 2011

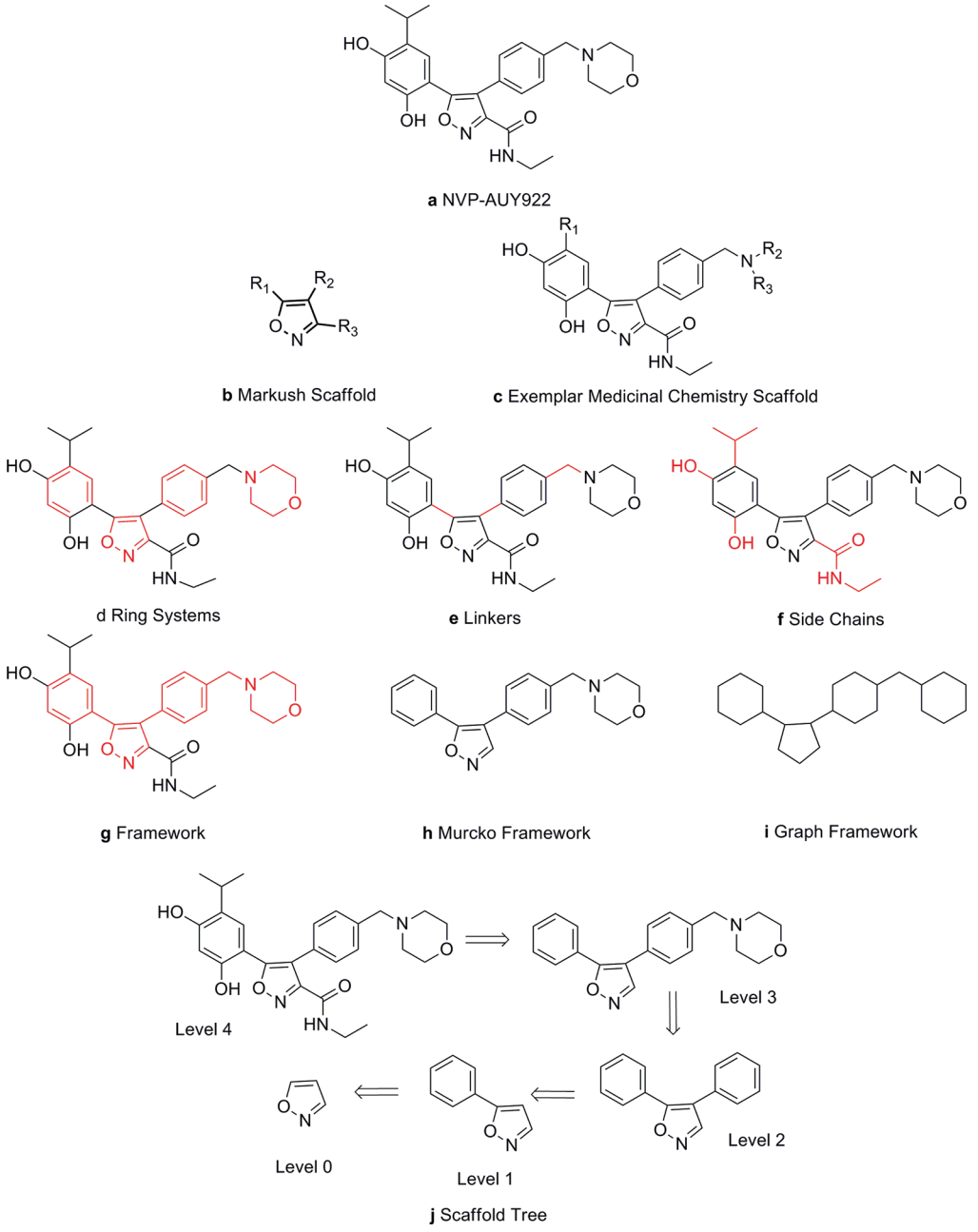



**Figure 2.** The HSP90 inhibitor NVP-AUY922 depicted using different scaffold representations.

based upon their chemical fingerprints and for each cluster the MCS is found: the compounds are then grouped based upon their MCS.[9] This method is data set dependent since different compound data sets will result in a different cluster assignment and therefore a different MCS.

The Murcko framework of a molecule can also be dissected into more than one ring system by cleaving linker bonds between rings in the Murcko framework. Compound libraries have been analyzed by the ring systems present,[10] which can be arranged in a hierarchical tree according to complexity.[11,12] The Scaffold Tree[13] is such an example of a hierarchical tree of ring systems. The Scaffold Tree methodology takes each molecule in a library and iteratively removes rings one by one, based on a set of prioritization rules, until only one ring remains. Each molecule has $n$+1 Levels numbered sequentially from Level 0 (the single remaining ring) up to Level $n$ (the whole molecule) where Level $n$-1 is the Murcko framework (Figure 2j). The scaffold hierarchies of each molecule are then combined into a tree for the whole library. The Scaffold Tree has been used in conjunction with biochemical activity data for structure−activity relationship (SAR) analysis[14−17] but, to our knowledge, has not yet been used for a scaffold diversity analysis of compound libraries, based upon the number and frequency of scaffolds present, as described herein.

A number of studies have been carried out to investigate the diversity of compound sets based upon the frequency of scaffolds present. For example, Bemis and Murcko identified 1179 scaffolds present in 5129 known drugs using the Murcko framework definition of a scaffold.[8] Half of the drugs in the data set were based on the 32 most frequently occurring scaffolds, suggesting


dx.doi.org/10.1021/ci2001428 |J. Chem. Inf. Model. 2011, 51, 2174–2185

that the scaffold diversity of known drugs is extremely low. A similar analysis of the CAS (Chemical Abstracts Service) Registry of over 24 million compounds also showed that a large percentage of organic compounds are represented by only a small percentage of scaffolds.[18] These findings suggest that the more frequently a scaffold has been used, the more likely it will be used again. A similar conclusion is drawn from a study that looks at the co-occurrence of fragments in the same molecule.[19] It was found that particular fragments and combinations of fragments were far more frequent than others and were therefore termed "Chemical Clichés".[19]

Compound libraries have also been analyzed in terms of the topology of the scaffolds present. One study describes ring systems present in the CAS Registry using three integer descriptors to represent the topology of the ring system, thereby allowing the ring systems to be plotted in 3-dimensional topology space.[20] Some areas of topology space were highly populated with significant voids also observed. A subset of the CAS Registry flagged for "Therapeutic Use" was mapped onto the topology space to represent medicinally relevant rings. Two bounds were found that contain these medicinally relevant rings, namely size and molecular complexity; indicating that the scaffold space of medicinally relevant rings is influenced by size constraints for druglike molecules and ease of synthesis. Similar conclusions were reached in a study that enumerated graph representations of scaffold topologies and examined frequency of occurrence in compound libraries.[21]

An analysis of the scaffolds present in a set of approximately 150,000 bioactive compounds found only 780 simple aromatic scaffolds.[22] A virtual library containing nearly 600,000 small aromatic scaffolds was produced to assess how the biologically active scaffolds covered chemical space. The scaffolds were clustered in a Self-Organizing Map (SOM) which demonstrated that biologically active scaffolds are sparsely distributed across the virtual library forming well-defined activity islands. The authors suggest three possible reasons for the lack of diversity of bioactive molecules:

1 Biological activity is limited to a small area of chemical space;
2 Most small aromatic scaffolds are synthetically inaccessible;
3 The chemical space of small aromatic scaffolds is so large that known bioactive compounds will only ever cover an insignificant proportion of this space.

A similar investigation[23] also suggested that some scaffolds are more popular due to the synthetic ease of attaching other medicinally relevant moieties; the accumulated synthetic and medicinal chemistry knowledge on these more popular scaffolds makes them more attractive for future use. A further study identified scaffolds that have selectivity for target-gene families but finds that these scaffolds are underrepresented in approved drugs.[24]

Metrics have been applied to quantify the distribution of molecules over scaffolds. For example NC50C and PC50C quantify the number of scaffolds and the percentage of scaffolds that represent 50% of molecules in a library.[23] Studies using these metrics again show that the distribution of molecules over scaffolds is skewed in compound libraries.[25] Shannon entropy may also be used to describe the distribution of molecules over scaffolds. A Shannon entropy of 0 indicates that all compounds contain the same scaffold; a high Shannon entropy indicates that each scaffold represents the same number of molecules and that the library is, therefore, evenly distributed over the represented scaffolds.[26]

The studies discussed above highlight the lack of scaffold diversity in many compound libraries. Recently Pitt et al. generated a collection of 24,847 virtual small aromatic rings named VEHICLe (virtual exploratory heterocyclic library).[27] Only 1701 of the VEHICLe ring systems were identified as synthesized (*i.e.* in existence). A machine learning approach predicted that over 3000 of the ring systems could easily be synthesized. This suggests that only a small area of scaffold space is covered by synthesized compounds and that a large area of scaffold space is synthetically accessible.

**Table 1. Summary of the Data Sets Used in the Scaffold Diversity Analysis**

| data set | description | compounds |
|---|---|---|
| ICRSC | compounds from the ICR's internal screening collection | 79,742 |
| VC | compounds from the ICR's preferred vendors | 1,923,627 |
| ICRFL | fragments from the ICR's internal screening collection | 2448 |
| CHEMBL | compounds from the EBI's ChEMBL database. The CHEMBL database compounds are taken from the medicinal chemistry literature.[31] | 530,038 |
| DBSM | small molecule drugs taken from DrugBank[34] | 4654 |
| DBAD | approved drugs taken from DrugBank[34] | 1361 |
| BIOFOC | compounds from BioFocus designed to target kinases[35] | 10,000 |

In this work we analyze the scaffold diversity of 7 representative compound libraries, including the ChEMBLdb, drug sets, vendor libraries, and in-house screening collections. We compare the Murcko framework and Scaffold Tree representations of scaffolds for the first time and show that Level 1 of the Scaffold Tree is useful for the characterization of scaffold diversity in compound libraries and offers advantages over the use of Murcko frameworks. This analysis also demonstrates that the majority of compounds in the libraries we analyzed contain only a small number of well represented scaffolds and that a high percentage of singleton scaffolds represents the remaining compounds. Tree Maps have recently been exemplified as a useful method for the two-dimensional (2D) depiction of structure activity relationships using dendrograms which incorporate molecule fragmentation hierarchies.[28] Here we demonstrate the use of Tree Maps to visualize the distribution of molecules over scaffolds, and the molecular similarity of the scaffolds, within a compound library. This novel use of Tree Maps provides easily interpretable depictions of compound library scaffold diversity for the medicinal chemistry community.

## METHODS

**Data Sets.** The scaffold diversity analyses were performed on 7 data sets containing both publicly available and proprietary compounds. They are described here and summarized in Table 1.

**ICR Screening Collection (ICRSC).** 79,742 Compounds from the Institute of Cancer Research (ICR) in-house hit discovery screening collection. This collection includes compounds selected from multiple commercial vendors based upon in-house developed filters for leadlike molecules as well as compounds synthesized in-house.

**Vendor Collection (VC).** 1,923,627 Compounds that are commercially available from 11 of the ICR's preferred vendors. Compound libraries were downloaded directly from the vendors, and compounds containing toxicophores were removed. Molecules with more than 35 heavy atoms or an AlogP greater than 6 were also removed.[29]

**ICR Fragment Library (ICRFL).** The fragment library (ICRFL) contains 2448 fragments, either synthesized at the ICR or purchased from vendors. [Fragment definition parameters: molecular weight: 150−300 Da (+20 Da for specific groups), AlogP: ≤ 3, H-bond acceptors: ≤ 5, H-bond donors: ≤ 3, topological polar surface area (TPSA): ≤ 75 Å$^2$, rotatable bonds: ≤ 4, heavy atoms: ≥ 10, rings: 1−3, ring size: 3−7 atoms, fused rings: ≤ 2, number of sulfur atoms: ≤ 1, number of halogen atoms: ≤ 1 (except fluorine), compounds containing toxicophores were removed.]

**ChEMBLdb (CHEMBL).** 530,038 Compounds from the EBI-ChEMBL database. ChEMBL-db consists of bioactive compounds taken from the medicinal chemistry literature and is manually curated by the EBI-ChEMBL team.[30] CHEMBL represents 80% of the version (08) of ChEMBLdb, Version 03 was used in this work.[31]

**DrugBank.** DrugBank is a bioinformatics and chemoinformatics resource that combines chemical drug data with target drug data. It contains small molecule drugs, FDA approved small molecule and biotech (protein/peptide) drugs, experimental drugs, and the protein or drug target sequences related to these drugs.[32,33] DrugBank Small Molecules (DBSM) contains 4654 small molecule drugs from DrugBank version 2.5, 1335 of which are also contained in DrugBank Approved Drugs (DBAD).[34] DrugBank Approved Drugs (DBAD) contains 1361 approved drugs from Drug Bank version 2.5.[34]

**BioFocus Kinase Focused Library (BIOFOC).** Library of 10,000 compounds from BioFocus designed to target kinases.[35]

**Scaffold Representations.** To analyze the scaffold diversity of the 7 data sets, two scaffold representations were used: Murcko frameworks[8] and the Scaffold Tree[13] both of which represent compounds containing cyclic systems. Murcko frameworks were generated in Pipeline Pilot 7.0[36] using the *Generate Fragments* component with the *FragmentsToGenerate* parameter set to *MurckoAssemblies*; all other parameters were kept as the default values.

The Scaffold Tree[13] is a hierarchical classification of chemical scaffolds. Murcko frameworks of the compounds in the library are generated; these are leaf nodes of the Scaffold Tree. Lower levels of the Scaffold Tree are obtained by iterative removal of rings according to a set of prioritization rules that are designed to retain the most richly functionalized ring systems and are intended to be intuitive to a synthetic medicinal chemist. This process continues until only one ring remains.[13] We have observed that, in the majority of cases, the most richly functionalized ring system is retained; however, in compounds containing an all carbon fused ring system, this scaffold is prioritized over a single ring heterocycle in the same molecule which could be regarded as more richly functionalized. The root node, the single remaining ring after fragmentation, is named Level 0, and subsequent levels or nodes in the tree are named numerically (Figure 2j). There can be any number of levels to the Scaffold Tree depending on the complexity of the molecules represented. Compounds of different complexity have different numbers of Levels in the Scaffold Tree; therefore, we sought a Level that the majority of compounds in our representative data set possess. All compounds of the Scaffold Tree possess Level 0; however, Level 0 always contains a single nonfused ring and is too generic to be a useful scaffold representation; for example, many molecules are reduced to the same single ring representation which lacks sufficient granularity. Higher levels of the Scaffold Tree were close or identical to the Murcko framework which often incorporates multiple ring systems. We observed fewer examples of concordance to the Murcko framework at Level 1 than at all other levels (excluding Level 0) although Level 1 scaffolds can be the same as the Murcko framework, for example, in the case of compounds defined as fragments. Level 2 scaffolds and above have a greater number of examples where the scaffold is the same as the Murcko framework. In addition, some small compounds (*e.g.* fragments) do not possess a Level 2 or above. We therefore used Level 1 of the Scaffold Tree in our analysis. In summary, the Scaffold Tree is a data set independent, rule based method, which is designed to retain the most richly functionalized ring systems. Level 1 scaffolds contain one or two rings, and as Level 1 is less complex than higher levels of the Scaffold Tree, the vast majority of compounds in our representative data sets possess a Level 1 scaffold; only single rings with no substituents are excluded from Level 1 (see below).

The Molecular Operating Environment (MOE) from the Chemical Computing Group[37] was used to generate the Scaffold Tree for each data set using the *linear fragmentation* function. An SVL script was applied to an SDF file; the *linear fragmentation* function was used to apply the Scaffold Tree fragmentation rules. The Level 1 scaffold of each compound is saved to a molecular database (.mdb file) along with the original molecule.

As mentioned above, both Level 1 scaffolds and Murcko frameworks can only represent molecules containing ring systems; therefore, acyclic molecules are omitted from the data analysis. This does not affect the ICRFL and BIOFOC data sets; for ICRSC, VC, and CHEMBL, between 0.06% and 2% of compounds are excluded. For DBSM and DBAD, 17.3% and 7.6% of molecules are excluded which include, for example, acyclic peptide drugs or development compounds. Molecules containing only a single ring with no substituents have one level, Level 0 of the Scaffold Tree, and are, therefore, omitted from the analysis. This only affects the VC, CHEMBL, and DBSM data sets where less than 0.15% of compounds are single rings with no substituents. One of the Scaffold Tree rules is the removal of 3-membered ring heterocycles (*e.g.* epoxides). The 3-membered ring is converted to a double bond. This step is carried out when side chains are removed from the molecule, before the iterative removal of other rings. As a consequence, compounds that contain only 3-membered heterocyclic rings are rendered acyclic before the iterative removal of rings and are therefore excluded from the analysis. This only affects the VC, CHEMBL, DBSM, and DBAD data sets where this rule applies to less than 0.5% of compounds.

**Scaffold Diversity Analysis.** In these analyses we investigate two types of diversity: the distribution of molecules over the unique scaffolds present in the data set and the structural diversity of these scaffolds. Of the methods described below, the scaffold counts and cumulative scaffold frequency plots provide information on the distribution of molecules over scaffolds, and the Tree Maps provide information on both the distribution and structural diversity.

**Scaffold Counts.** The scaffold diversity analysis was performed on the ICRSC, VC, ICRFL, CHEMBL, DBSM, DBAD, and BIOFOC data sets. For each data set the Murcko frameworks and Level 1 of the Scaffold Tree were defined for each compound

and the following steps performed using Pipeline Pilot.[36] The numbers of unique Murcko frameworks and Level 1 scaffolds for each data set were counted, along with the number of molecules they represent; this is referred to as the scaffold frequency. The number of singleton scaffolds was also recorded; singleton scaffolds are scaffolds that are only present in one exemplar molecule.

**Cumulative Scaffold Frequency Plots (CSFP).** Scaffold frequency is the number of molecules that contain a particular scaffold; the scaffold frequency can be represented as a percentage of total molecules in the data set. To generate cumulative scaffold frequency plots (CSFP), the scaffolds are sorted by their scaffold frequency (most frequent to least frequent) the cumulative percentage of scaffolds is then plotted against the cumulative scaffold frequency as a percentage of total molecules. CSFPs were generated for each data set using both the Murcko and Level 1 scaffold representations. From the cumulative frequency plots, the percentage of scaffolds that represent $n$ percent of compounds can be determined $P_n$. For example $P_{50}$ is the percentage of scaffolds that represent 50% of compounds; this measure has been used in various scaffold diversity analyses and is often termed PC50C.[23] The ratio of scaffolds to compounds ($N/M$) and the ratio of singleton scaffolds to all scaffolds ($N_s/N$) are also used to assess the diversity of scaffold space.

**Tree Maps.** Tree Maps are visualizations of hierarchical data structures and were introduced in 1992 by Shneiderman to visualize the directory tree of hard disks.[38] Tree Maps use a 2D space-filling approach where rectangles or circles represent each leaf of a hierarchical tree. The size and color of each rectangle or circle can correspond to specific properties of the data being represented. Tree Maps have been used previously to visualize hierarchical clusters of compounds and their biological activity data.[28,39] Rather than visualizing the whole hierarchical Scaffold Tree with Tree Maps we have visualized all Level 1 scaffolds present in each data set and have clustered the scaffolds based on their structural similarity. We have used circular Tree Maps rather than rectangular Tree Maps to better highlight clusters of scaffolds. The color and area of the circles represents the scaffold frequency of the scaffolds they represent. This allows visualization of both scaffold structural diversity and the distribution of compounds over scaffolds. To our knowledge this is the first application of Tree Maps to visualize the distribution and chemical diversity of compound libraries.

Tree Maps were created using the software TreeMap from MacroFocus.[40] First, the Level 1 scaffolds of each data set were clustered using FCFP_2 fingerprints. The *Cluster Molecules* component in Pipeline Pilot was applied with the average number of compounds *per* cluster set to 20. This component selects a molecule from the data set at random as the first cluster center and then selects the remaining cluster centers to give maximum dissimilarity to the first cluster center and each other. The remaining molecules are then assigned to each cluster based upon their similarity to the cluster center. This method is order dependent, as the random molecule selection is dependent on the order the molecules enter the component. As the *Cluster Molecules* component presorts the data we used the *Cluster Data* component to test the order dependency of the clustering algorithm. The clustering protocol was applied 5 times; for each run, a random number was generated for each compound using the current time in 24-h format as the seed for the *Random Number* component. The compounds were then sorted by the random number and sent to the *Cluster Data* component. Thus, for each run, the compounds are entering the *Cluster Data* component in a different order. For each run of the protocol the mean Kelley spread and distance of the clusters were calculated.[41,42] The spread is based on the mean pairwise similarity of the members of a cluster, and the distance reflects how dissimilar one cluster is from another. The mean spread and distance for each of the 5 runs were consistent, within one standard deviation, when tested with all 7 data sets. In summary, the order dependency of the *Cluster Molecules* component did not have a major effect on the clusters used to visualize the data sets in the Tree Maps.

After clustering, each scaffold had a cluster number attributed to it which could be used in the TreeMap software to group the scaffolds based on the clusters to which they belong. The scaffold frequency attributed to each scaffold was also used in the visualization.

In the Tree Maps, scaffolds are represented by circles where the area of the circle is proportional to the scaffold frequency. The color of the circle is also related to the scaffold frequency. The scaffolds are grouped into gray circles, which represent the cluster to which they belong. The distances between clusters and scaffolds in the Tree Maps are not representative of the structural or property similarity of clusters and scaffolds.

**Table 2. Murcko Framework Analysis: Results of the Scaffold Diversity Analysis on the ICRSC, VC, ICRFL, CHEMBL, DBSM, DBAD, and BIOFOC Data Sets Using Murcko Frameworks[a]**

| data set | $M$ | $N$ | $N_s$ | $N/M$ | $N_s/N$ | $P_{25}$ | $P_{50}$ | $P_{75}$ |
|---|---|---|---|---|---|---|---|---|
| ICRSC | 76,563 | 33,050 | 23,123 | 0.41 | 0.70 | 1.23 | 9.93 | 39.8 |
| VC | 1,922,434 | 388,952 | 237,617 | 0.20 | 0.61 | 0.22 | 2.01 | 15.3 |
| ICRFL | 2448 | 1146 | 822 | 0.47 | 0.72 | 1.67 | 12.0 | 46.6 |
| CHEMBL | 519,362 | 153,199 | 102,548 | 0.29 | 0.67 | 0.49 | 4.49 | 24.1 |
| DBSM | 3849 | 2061 | 1680 | 0.54 | 0.82 | 1.05 | 12.6 | 53.3 |
| DBAD | 1258 | 785 | 633 | 0.57 | 0.81 | 3.58 | 19.9 | 59.9 |
| BIOFOC | 10,000 | 2498 | 1559 | 0.25 | 0.62 | 0.80 | 5.36 | 20.9 |

[a] $M$ = number of compounds, $N$ = number of Murcko frameworks, $N_s$ = number of singleton scaffolds, $P_n$ = percentage of scaffolds that represent $n$ percent of compounds.

## RESULTS AND DISCUSSION

**Scaffold Counts.** Tables 2 and 3 show the number of compounds included in the analysis ($M$) and the number of Murcko frameworks and Level 1 scaffolds present in each data set ($N$) as well as the number of singleton scaffolds ($N_s$). The ratios of scaffolds to molecules ($N/M$) and singleton scaffolds to total scaffolds ($N_s/N$) are also reported. $P_n$ values ($P_{25}$, $P_{50}$, and $P_{75}$ in Tables 2 and 3) indicate the percentage of scaffolds that represent $n$ percent of compounds; thus $P_{75}$ is the percentage of scaffolds that represent 75% of all compounds in the data set. The significance of these figures is discussed alongside the cumulative scaffold frequency plots below.

When analyzed using Level 1 scaffolds, the BIOFOC and VC data sets have an extremely low ratio of scaffolds to molecules ($N/M$ = 0.02 and 0.04, respectively) indicating that these data sets contain heavily represented scaffolds. The CHEMBL and ICRSC data sets also have a low proportion of scaffolds ($N/M$ = 0.13 and 0.16, respectively). The DBAD, DBSM, and ICRFL

**Table 3. Scaffold Tree Analysis: Results of the Scaffold Diversity Analysis on the ICRSC, VC, ICRFL, CHEMBL, DBSM, DBAD, and BIOFOC Data Sets Using Level 1 of the Scaffold Tree[a]**

| data set | $M$ | $N$ | $N_s$ | $N/M$ | $N_s/N$ | $P_{25}$ | $P_{50}$ | $P_{75}$ |
|---|---|---|---|---|---|---|---|---|
| ICRSC | 79,563 | 12,520 | 8637 | 0.16 | 0.69 | 0.22 | 1.34 | 7.70 |
| VC | 1,922,433 | 81,368 | 62,889 | 0.04 | 0.77 | 0.026 | 0.14 | 0.70 |
| ICRFL | 2448 | 1074 | 792 | 0.43 | 0.74 | 2.04 | 10.1 | 43.0 |
| CHEMBL | 519,341 | 68,370 | 53,385 | 0.13 | 0.78 | 0.032 | 0.35 | 3.06 |
| DBSM | 3843 | 2012 | 1668 | 0.52 | 0.83 | 1.15 | 10.8 | 52.3 |
| DBAD | 1253 | 691 | 537 | 0.55 | 0.78 | 2.93 | 15.8 | 54.7 |
| BIOFOC | 10,000 | 167 | 49 | 0.02 | 0.35 | 1.00 | 3.10 | 8.34 |

[a] $M$ = number of compounds, $N$ = number of Level 1 scaffolds, $N_s$ = number of singleton scaffolds, $P_n$ = percentage of scaffolds that represent $n$ percent of compounds.

data sets ($N/M$ = 0.55, 0.52, and 0.43, respectively) have close to one scaffold for every two molecules suggesting they are the most scaffold diverse data sets. The ratio of scaffolds to molecules ($N/M$) should be used in conjunction with the number of singleton scaffolds to provide accurate information on the distribution of molecules over scaffolds. For example, DBSM has $N/M$ = 0.52 (*i.e.* there are 1.9 molecules to every scaffold) suggesting that the data set is scaffold diverse. However, the proportion of singleton scaffolds to scaffolds ($N_s/N$) is 0.83; therefore, 83% of scaffolds (1670 scaffolds) represent only 1 molecule each, and 17% of scaffolds (342 scaffolds) represent the remaining 2173 molecules (an average of 6.3 molecules *per* scaffold). Table 3 indicates that, in most cases, a large proportion of the Level 1 scaffolds are singletons ($N_s/N$ > 0.6), suggesting that the distribution of molecules over scaffolds is uneven. An exception is the BIOFOC data set ($N_s/N$ = 0.35); despite having the lowest proportion of scaffolds ($N/M$ = 0.02), a low proportion of these are singletons. In the case of the BIOFOC data set the molecules are more evenly distributed over a small number of scaffolds. This result is expected since the BIOFOC data set was designed as a screening collection containing a selection of kinase inhibitor scaffolds that are equally represented.[35]

The Murcko framework analysis delivers similar overall conclusions (Table 2). BIOFOC and VC have the lowest proportion of scaffolds ($N/M$ = 0.25 and 0.20, respectively), and DBAD, DBSM, and ICRFL have the highest proportion of scaffolds ($N/M$ = 0.57, 0.54, and 0.47, respectively). A high proportion of scaffolds are singletons ($N_s/N$ > 0.6), and $N_s/N$ values are more uniform across the 7 data sets than for the Level 1 scaffold analysis. One difference in the analyses using Murcko frameworks and Level 1 scaffolds is the proportions of scaffolds to molecules ($N/M$). The lowest proportion of scaffolds when using the Murcko framework representations (for VC $N/M$ = 0.20) is higher than the lowest proportion of scaffolds when using Level 1 scaffold representations (for BIOFOC $N/M$ = 0.02). The range of proportions is also narrower for Murcko frameworks (range of $N/M$ = 0.37) compared to Level 1 scaffolds (range of $N/M$ = 0.53). Murcko frameworks are a more discriminating representation of a scaffold (*i.e.* they carry greater chemical description) than Level 1 Scaffolds, and, therefore, more unique scaffolds are defined by Murcko frameworks. Thus, there is a higher proportion of Murcko scaffolds present in data sets. The difference in $N/M$ for the Murcko and Level 1 analyses are significant for most cases examined apart from ICRFL, DBSM, and DBAD where the difference is very small (Table 4). A possible reason is that these three data sets contain a relatively high percentage of molecules which have one ring (12.5−19.4%) as defined in Pipeline Pilot[36] (Table 4). In the case of molecules containing a single ring, Level 0 of the Scaffold Tree is this single ring and is also the Murcko framework, whereas the next Level up the hierarchical Scaffold Tree (Level 1) is the single ring plus its substituents (the whole molecule). For such compounds, the Level 1 scaffold is more descriptive than the Murcko framework, and, therefore, more unique Level 1 scaffolds than Murcko frameworks are generated. The ICRFL, DBSM, and DBAD data sets have a much higher proportion of fragmentlike compounds containing only one ring than other data sets in our analysis and could, we propose, explain why the results for the Murcko and Level 1 analyses for these data sets are more similar.

**Table 4. Percentage of Compounds in Each Data Set That Have Less than One Ring**

| data set | number of rings = 1 | $N/M$ (Murcko) | $N/M$ (Level 1) |
|---|---|---|---|
| ICRSC | 5.0% | 0.41 | 0.16 |
| VC | 2.5% | 0.20 | 0.04 |
| ICRFL | 12.5% | 0.47 | 0.43 |
| CHEMBL | 6.8% | 0.29 | 0.13 |
| DBSM | 19.4% | 0.54 | 0.52 |
| DBAD | 15.6% | 0.57 | 0.55 |
| BIOFOC | 0.1% | 0.25 | 0.02 |

In summary, we have shown that the proportion of scaffolds present in a data set ($N/M$), in conjunction with the proportion of singleton scaffolds ($N_s/N$), is a useful indicator of scaffold diversity across a diverse range of compound libraries. In most of the data sets tested, molecules are unevenly distributed, with a small number of highly populated scaffolds and a large number of singletons. Murcko frameworks are a more discriminating representation of a scaffold than Level 1 scaffolds, and, as a result, more unique scaffolds are defined by Murcko frameworks. For molecules containing one ring, Level 1 scaffolds provide a more granular representation in comparison to Murcko scaffolds.

**Cumulative Scaffold Frequency Plots (CSFP) and $P_n$ Values.** The CSFPs are shown in Figure 3 for the Level 1 scaffolds and Figure 4 for the Murcko frameworks. The CSFPs give an indication of the distribution of molecules over scaffolds. In the extreme case where each scaffold represents the same number of compounds, the plot would be diagonal from (0%, 0%) to (100%, 100%); therefore, the closer the curve is to the diagonal, the more evenly distributed the data set.

In the case of Level 1 scaffolds (Figure 3) the CSFPs give similar conclusions to the proportion of scaffolds to molecules as shown in Table 3. DBAD, DBSM, and ICRFL are closest to the diagonal, indicative of a more even distribution. The BIOFOC, VC, CHEMBL, and ICRSC data sets are furthest from the diagonal and are least evenly distributed. All curves begin with a very steep gradient; this indicates the presence of scaffolds that represent a large proportion of the data set. The shallow region of the curve represents the high proportion of singleton scaffolds. An interesting example is the BIOFOC data set, which has a much lower ratio of singleton scaffolds to overall scaffolds ($N_s/N$ = 0.35) despite having a low proportion of scaffolds overall ($N/M$ = 0.02). This profile is represented by a lower gradient early in the plot compared to other data sets (VC, CHEMBL, and ICRSC, Tables 2 and 3). The BIOFOC curve

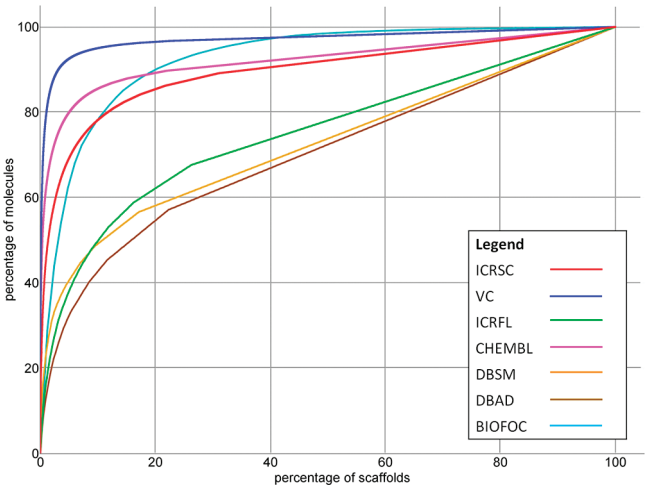



**Figure 3.** Scaffold Tree analysis: cumulative scaffold frequency plot showing the distribution of compounds over Level 1 scaffolds in the ICRSC, VC, ICRFL, CHEMBL, DBSM, DBAD, and BIOFOC data sets.

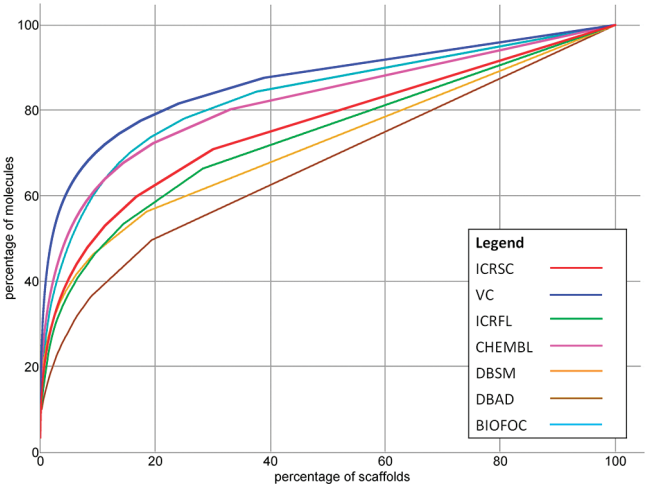



**Figure 4.** Murcko framework analysis: cumulative scaffold frequency plot showing the distribution of compounds over Murcko framework scaffolds in the ICRSC, VC, ICRFL, CHEMBL, DBSM, DBAD, and BIOFOC data sets.

also levels off later than for the other data sets due to the lower proportion of singletons in the data set.

The CSFP for the Murcko framework representations (Figure 4) indicates similar overall trends to the CSFP analysis using Level 1 scaffolds, *i.e.* the DBAD, DBSM, and ICRFL data sets are more evenly distributed than the VC, ICRSC, CHEMBL, and BIOFOC data sets. However, the Murcko framework CSFP analysis is less discriminatory between data sets. We propose that this is a result of the more granular Murcko scaffold definition (i.e. they carry greater chemical description) which enhances the apparent scaffold diversity with respect to use of Level 1 scaffolds.

The information obtained from the CSFPs can be quantified using $P_n$ values; this is the percentage of scaffolds that represent $n$ percent of compounds. These values are shown in Tables 2 and 3. The $P_n$ values reflect the conclusions discussed above. For example VC has a low proportion of Level 1 scaffolds and is unevenly distributed in the CSFP; its $P_{25}$, $P_{50}$, and $P_{75}$ values are

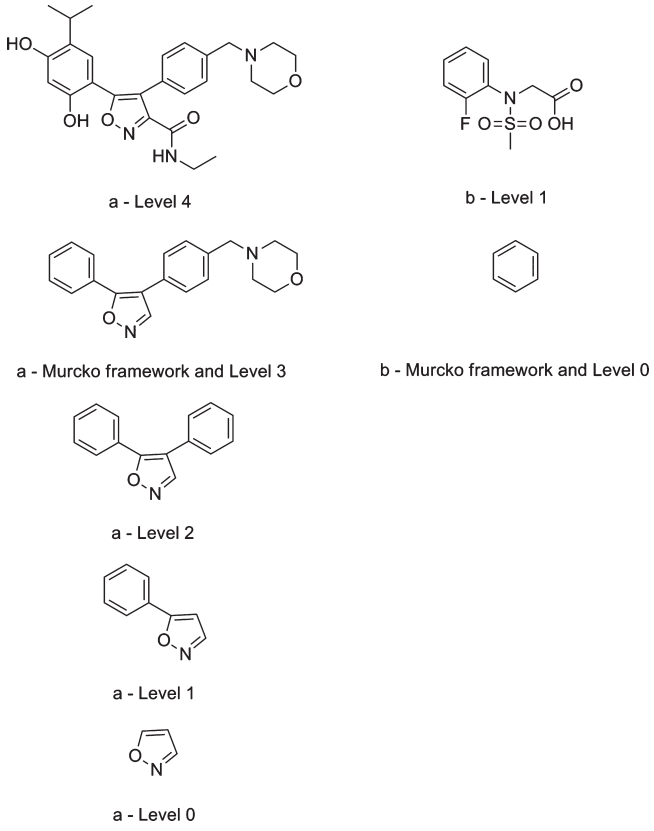



**Figure 5.** Examples of how compounds of different complexity are represented by Murcko frameworks and Level 1 scaffolds. Each molecule had $n$+1 Levels numbered sequentially from Level 0 (the single remaining ring) up to Level $n$ (the whole molecule) where Level $n$-1 is the Murcko framework: compound a: A typical leadlike/druglike chemical structure; compound b: A typical fragmentlike chemical structure.

$2.50 \times 10^{-2}$, 0.13, and 0.70, respectively. This indicates that 75% of the data set is represented by 0.70% of unique scaffolds and confirms that the data set is unevenly distributed with the majority of compounds represented by an extremely small proportion of scaffolds. The BIOFOC data set has the lowest proportion of scaffolds of all the data sets, but the proportion of singletons and the CSFP analysis indicates that molecules are distributed more evenly over these few scaffolds. The $P_{25}$, $P_{50}$, and $P_{75}$ for BIOFOC are 1.00, 3.10, and 8.34, respectively. These values are higher than those for other data sets with low $M/N$ supporting our evidence that BIOFOC is more evenly distributed over scaffolds.

**Murcko Frameworks vs Level 1 Scaffold Tree Analyses.** From these analyses several differences are apparent in the use of Level 1 scaffolds and Murcko frameworks to characterize the scaffold diversity of chemical libraries. Murcko frameworks deliver a more even distribution of compounds over scaffolds. We propose that this is because Murcko frameworks are more granular in definition (i.e. they carry greater structural description) such that there are more unique scaffolds in a data set. This can be a drawback, for example, larger molecules with many ring systems will likely have a Murcko framework similar to the original molecule which does not represent the molecular core (Figure 5a). However, for libraries of fragmentlike molecules containing only one ring, many compounds will be represented


dx.doi.org/10.1021/ci2001428 |J. Chem. Inf. Model. 2011, 51, 2174–2185

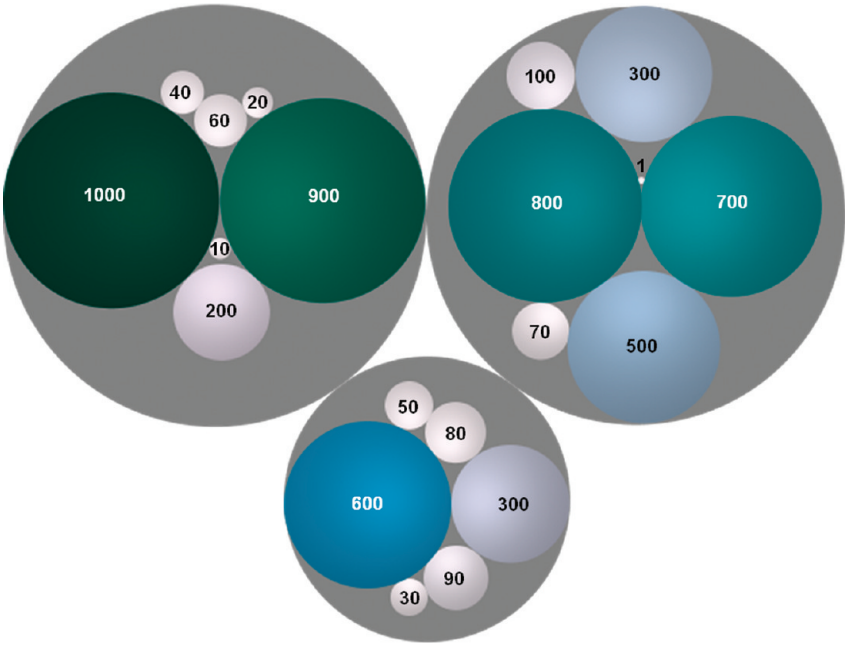



**Figure 6.** Example Tree Map. The colored circles represent scaffolds and are labeled with their scaffold frequency. The area and color of the circles relate to the scaffold frequency. Scaffold circles are grouped into gray circles if the scaffolds are in the same cluster.

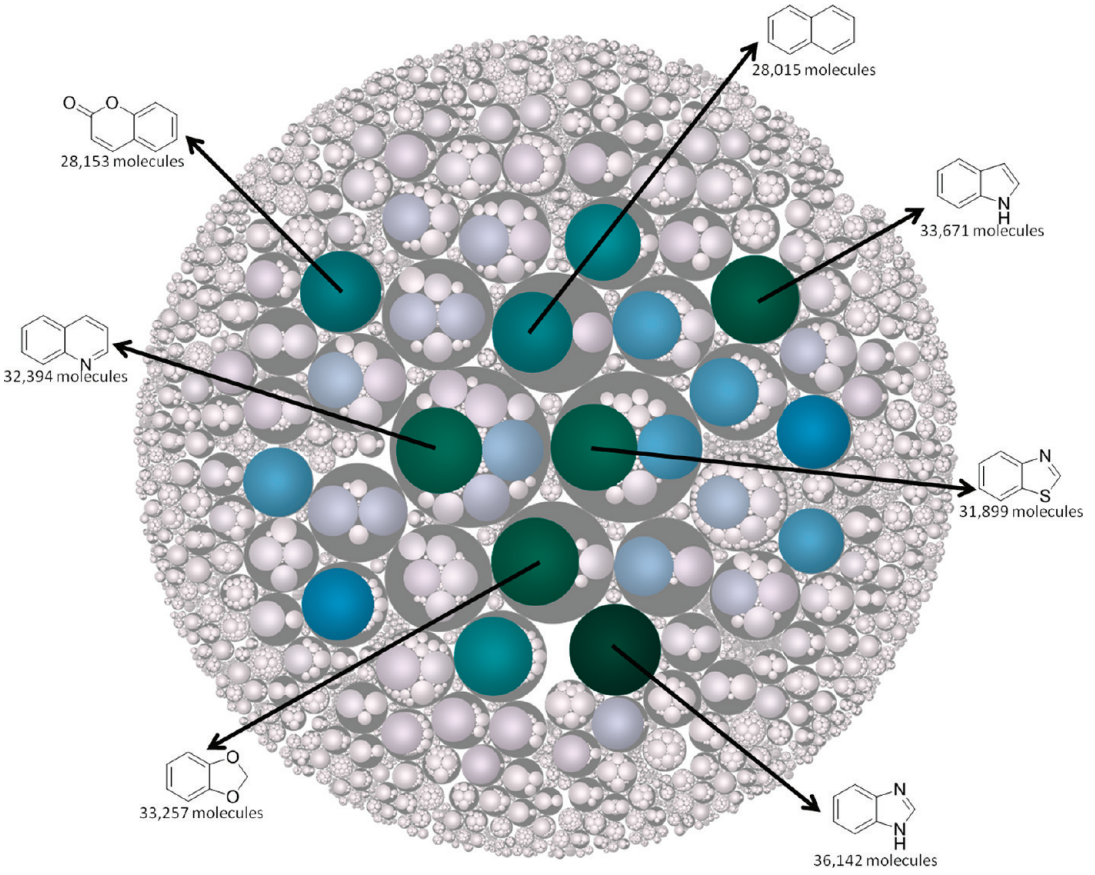



**Figure 7.** Tree Map of the VC data set Level 1 scaffolds. Scaffolds are represented by colored circles, the area and color of the circles relate to the scaffold frequency, gray circles represent clusters of scaffolds. Tree Maps illustrate the large proportion of singleton scaffolds in the data sets (many small white circles) and the presence of highly populated scaffolds (few large green circles).


dx.doi.org/10.1021/ci2001428 |*J. Chem. Inf. Model.* 2011, 51, 2174–2185

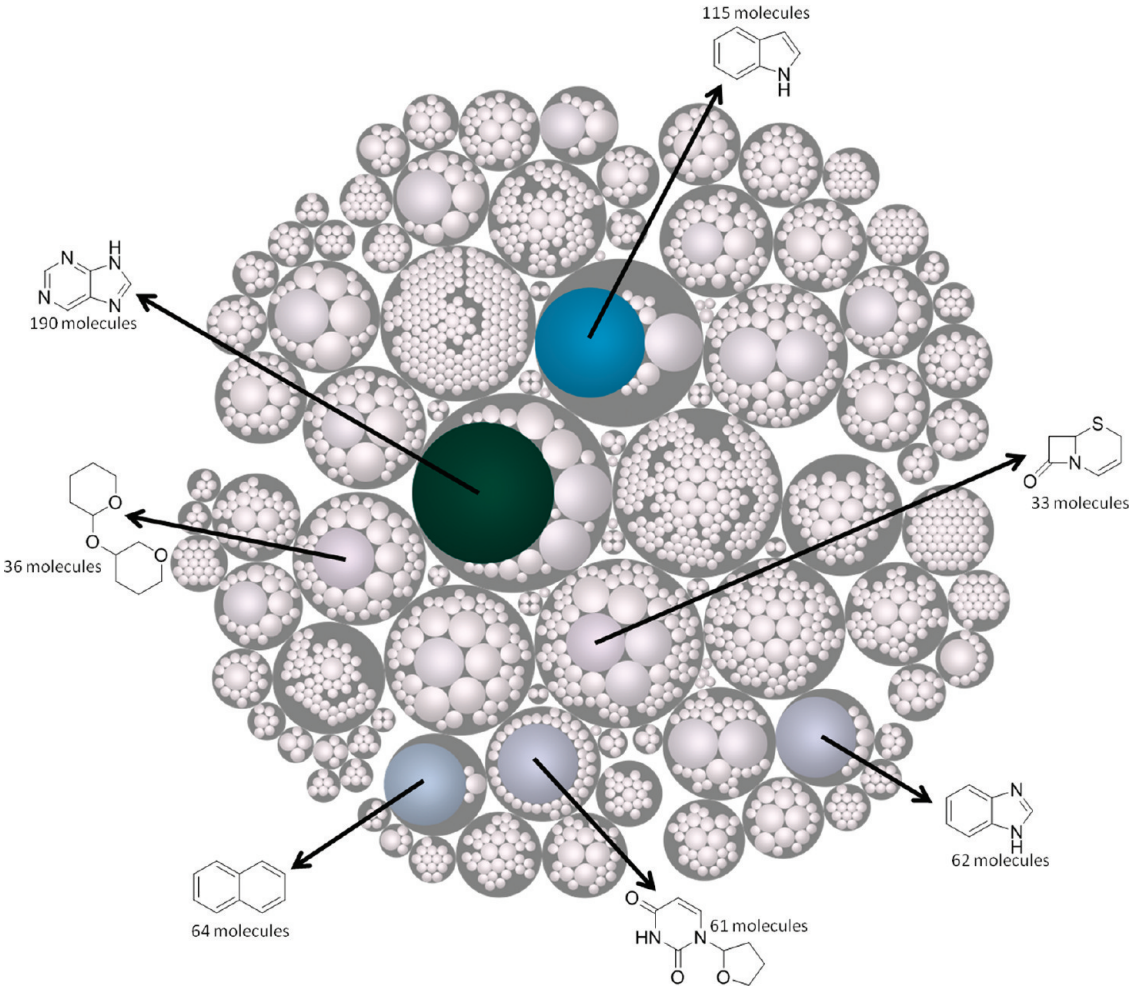



**Figure 8.** Tree Map of the DBSM data set Level 1 scaffolds. Scaffolds are represented by colored circles, the area and color of the circles relate to the scaffold frequency, gray circles represent clusters of scaffolds. Tree Maps illustrate the large proportion of singleton scaffolds in the data sets (many small white circles) and the presence of highly populated scaffolds (few large green circles).

by a single common scaffold in the Murcko framework (Figure 5b). Level 1 scaffolds perform better in both these scenarios. Larger compounds are reduced to two ring scaffolds, which better represent the core of the molecule (Figure 5a), while fragments with one ring retain greater structural information (Figure 5b). In addition, we show that Level 1 scaffolds better highlight the separation between more and less scaffold diverse data sets. We therefore used Level 1 scaffolds for our further analyses.

**Tree Maps.** We have presented the scaffold diversity of compound libraries using the distribution of molecules over scaffolds. This is different from structural diversity, a term which we use here to describe differences in overall chemical structure. Our analysis shows that the compound libraries we studied are unevenly distributed over scaffolds; however, the well represented scaffolds may be structurally diverse. To examine this aspect of scaffold diversity we visualize the structural similarity of scaffolds using Tree Maps.[28,38]

The Level 1 scaffolds for each data set were clustered by their fingerprint similarity using FCFP_2 fingerprints. The scaffolds were then visualized using Tree Maps.[38] Figure 6 shows a simple example of a Tree Map, each colored circle represents a scaffold, the scaffold circles are grouped into gray circles which represent the cluster to which the scaffolds belong. The area of each scaffold circle is proportional to the scaffold frequency, the largest circles have the highest scaffold frequency, and the smallest circles have the lowest scaffold frequency. The color of the circle is also related to the scaffold frequency. In the example each circle is labeled with the scaffold frequency to illustrate how the size and color of the circles relate to scaffold frequency, although the actual Tree Maps are too complex to show these numerical labels.

Tree Map representations of the Level 1 scaffolds for the chemical libraries under study are depicted in Figures 7−9 and S1−S4. Tree Maps illustrate the large proportion of singleton scaffolds in the data sets (many small white circles) and the presence of highly populated scaffolds (few large green circles). The added information of the scaffold clusters better depicts the structural diversity of the highly populated scaffolds. For example, the BIOFOC data set is designed to contain kinase inhibitor-like compounds, and the hinge binding scaffold is often highly conserved in inhibitors of this gene family. This is illustrated in the Tree Map for the BIOFOC data set (Figure 9) where the most highly populated scaffolds are clustered together. The DBAD and DBSM data sets (Figures S4 and 8) are more diverse, consistent with our previous analyses, here scaffolds are more evenly represented, and the most popular scaffolds are more structurally diverse and therefore found in different clusters. The ICRSC and VC (Figures S1 and 7) data sets are unevenly distributed

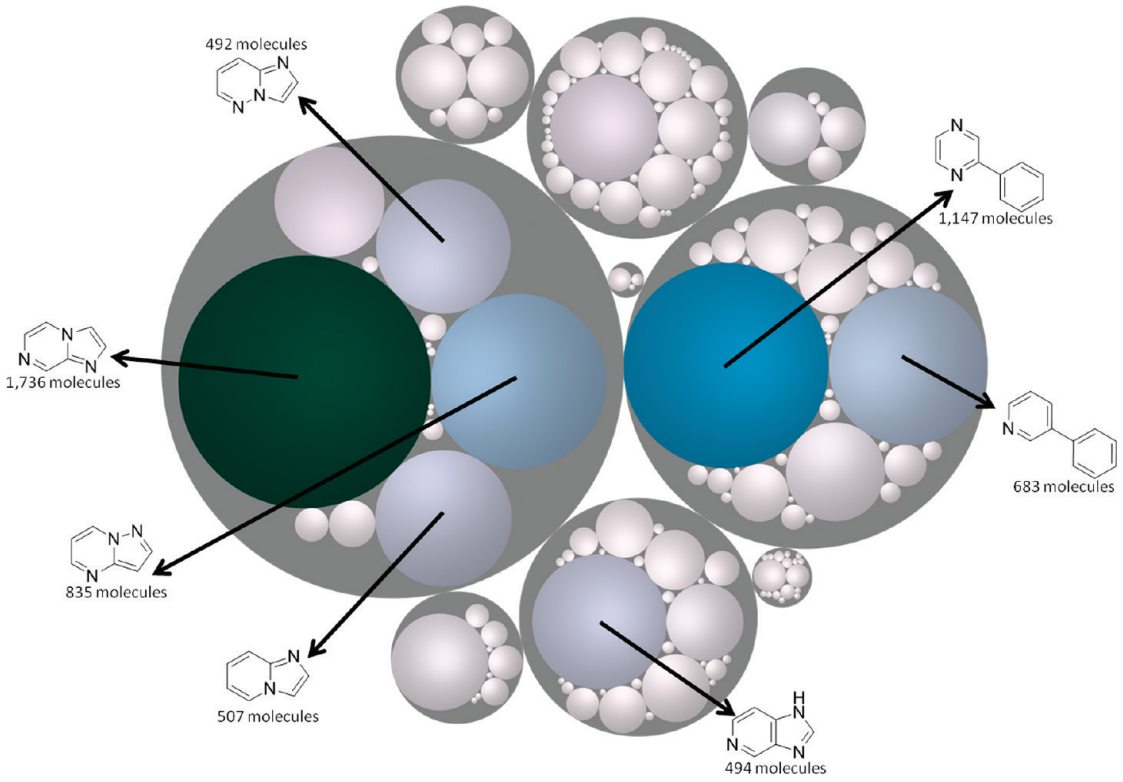



**Figure 9.** Tree Map of the BIOFOC data set Level 1 scaffolds. Scaffolds are represented by colored circles, the area and color of the circles relate to the scaffold frequency, gray circles represent clusters of scaffolds. Tree Maps illustrate the large proportion of singleton scaffolds in the data sets (many small white circles) and the presence of highly populated scaffolds (few large green circles).

over scaffold space; however, the highly populated scaffolds are in different clusters of the Tree Map indicative of structural diversity within the overall library.

In summary, we have used Tree Maps to visualize the diversity of compound libraries. We show that Tree Maps are an effective way to illustrate both the distribution and structural diversity of chemical scaffolds. The Tree Map visualization of the compound libraries under study clearly shows the presence of highly populated scaffolds as well as singleton scaffolds and illustrates the structural similarity of scaffold space.

## CONCLUSIONS

We have shown that a representative set of compound libraries commonly used in drug discovery are predominantly distributed over a small number of highly populated scaffolds with a concomitantly high number of singleton scaffolds when analyzed using both Murcko frameworks and Level 1 scaffold definitions. High representation in small areas of scaffold space is useful in libraries focused on particular biological target classes; for example the BIOFOC kinase focused library, where dense coverage of pharmacophore space is desired. However, this dense coverage may also represent significant redundancy in screening collections due to over population with structurally similar compounds. Poorly represented or singleton scaffolds may also be problematic in screening collections; for example, hit confirmation for molecules that are single exemplars of a scaffold can be hampered by the paucity of close analogs available for screening. In addition, it is often difficult to readily produce SAR data through rapid screening of close analogs. Thus screening collections should ideally be diversified by inclusion of more representative exemplars of singleton scaffolds.

We found that Level 1 scaffolds better highlight the differences between compound data sets than Murcko frameworks. Analysis using Murcko frameworks shows that data sets are more evenly distributed over scaffolds then analysis using Level 1 scaffolds. We propose that this difference arises because Murcko frameworks are a more granular representation of a molecule than Level 1 scaffolds (by virtue of the fact that Murcko frameworks incorporate substituents), and, therefore, data sets appear to contain more unique scaffolds, and appear more diverse, than the analysis by Level 1 scaffolds described here. Level 1 scaffolds are also applicable across a wider range of molecular weight and complexity than Murcko frameworks; for example, they encompass substituents present on single ring fragmentlike molecules which are increasingly important and prevalent constituents of compound libraries. Level 1 scaffolds provide a useful compromise between the minimalistic Level 0 single ring scaffold representation and the more granular Murcko definition. For these reasons we propose that Level 1 scaffolds are better suited to the analysis and cross comparison of diverse compound libraries ranging from fragmentlike and leadlike to Rule of Five compliant.

For the first time, we have used Tree Maps to visualize compound library composition and show that they are an effective way of illustrating both the distribution and structural diversity of chemical scaffolds. Tree Map visualization of the compound libraries under study clearly shows the presence of highly populated scaffolds as well as singleton scaffolds. In addition, Tree Maps clearly illustrate the structural similarity of constituent scaffolds. Tree Maps therefore provide a useful tool for medicinal chemists to assess the scaffold diversity of screening

libraries, assisting the prioritization of synthetic efforts directed toward library diversification. We are currently developing other methods for effectively visualizing and comparing scaffold distribution and diversity for the analysis and design of compound libraries for hit generation.

## ■ ASSOCIATED CONTENT

**Supporting Information.** Tree Map representations of the ICRSC, ICRFL, CHEMBL, and DBAD data set Level 1 scaffolds are shown in Figures S1−S4, respectively. This material is available free of charge via the Internet at http://pubs.acs.org.

## ■ AUTHOR INFORMATION

**Corresponding Author**

*E-mail: nathan.brown@icr.ac.uk (N.B.), julian.blagg@icr.ac.uk (J.B.).

## ■ ACKNOWLEDGMENT

Sarah Langdon is funded by the Institute of Cancer Research; Nathan Brown and Julian Blagg are funded by Cancer Research UK Grant No. C309/A8274.

We would like to thank the following: Guido Kirsten from the Chemical Computing Group for providing the SVL script used to produce the Scaffold Tree and Mike Cherry and Willem van Hoorn from Accelrys for their valuable discussion and solutions to the order dependency of clustering in Pipeline Pilot. We would also like to thank Berry Matijssen, Caterina Barillari, Ian Collins, Swen Hoelder, and Bissan Al-Lazikani for helpful discussions.

## ■ REFERENCES

(1) Villar, H. O.; Hansen, M. R. Design of chemical libraries for screening. *Expert Opin. Drug Discovery* **2009**, *4*, 1215–1220.

(2) Akritopoulou-Zane, I.; Hajduk, P. J. Kinase-targeted libraries: The design and synthesis of novel, potent, and selective kinase inhibitors. *Drug Discovery Today* **2009**, *14*, 291–297.

(3) Markush, E. A. Pyrazolone dye and process of making the same, Pharma Chemical Corp, Patent Number: 1506316, United States. 1924. The patent describes the Markush structure as "The yellow coloring matter which may be obtained by coupling to halogen-substitution products of pyrazolone, a diazotized unsulphonated material selected from the group consisting of aniline, homologues of aniline and halogen substitution products of aniline".

(4) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*, Revised ed.; Springer: Dordrecht, 2007.

(5) Brough, P. A; Aherne, A.; Barril, X.; Borgognoni, J.; Boxall, K.; Cansfield, J. E.; Cheung, K.-M. J.; Collins, I.; Davies, N. G. M.; Drysdale, M. J.; Dymock, B.; Eccles, S. A.; Finich, H.; Fink, A.; Hayes, A; Howes, R.; Hubbard, R. E.; James, K.; Jordan, A. M.; Lockie, A.; Martins, V.; Massey, A.; Matthews, T. P.; McDonald, E.; Northfield, C. J.; Pearl, L. H.; Prodromou, C.; Ray, S.; Raynaud, F. I.; Roughley, S. D.; Sharp, S. Y.; Surgenor, A.; Walmsley, D. L.; Webb, P.; Wood, M.; Workman, P.; Wright, L. 4, 5-Diarylisoxazole Hsp90 Chaperone Inhibitors: Potential Therapeutic Agents for the Treatment of Cancer. *J. Med. Chem.* **2008**, *51*, 196–218.

(6) Drysdale, M. J.; Dymock, B. M.; Finch, B.; Webb, P.; McDonald, E.; James, K. E.; Cheung, K.; Matthews, T. *Isoxazole Compounds as Inhibitors of Heat Shock Proteins*, Patent Number: US 2006/0241106 A1, United States, 2006.

(7) Langdon, S. R.; Ertl, P.; Brown, N. Bioisosteric Replacement and Scaffold Hopping in Lead Generation and Optimization. *Mol. Inf.* **2010**, *29*, 366–385.

(8) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(9) Nicolaou, C. A.; Tamura, S. Y.; Kelley, B. P.; Bassett, S. I.; Nutt, R. F. Analysis of Large Screening Data Sets via Adaptively Grown Phylogenetic-Like Trees. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1069–1079.

(10) Lewell, X. Q.; Jones, A. C.; Bruce, C. L.; Harper, G.; Jones, M. M.; Mclay, I. M.; Bradshaw, J. Drug Rings Database with Web Interface. A Tool for Identifying Alternative Chemical Rings in Lead Discovery Programs. *J. Med. Chem.* **2003**, *46*, 3257–3274.

(11) Kho, R.; Hodges, J. A.; Hansen, M. R.; Villar, H. O. Ring Systems in Mutagenicity Databases. *J. Med. Chem.* **2005**, *48*, 6671–6678.

(12) Wilkes, S. J.; Janes, J.; Su, A. I. HierS: Hierarchical Scaffold Clustering Using Topological Chemical Graphs. *J. Med. Chem.* **2005**, *48*, 3182–3193.

(13) 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.

(14) Renner, S.; van Otterlo, W. A. L.; Seoane, M. D.; 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.

(15) Clark, A. M.; Labute, P. Detection and Assignment of Common Scaffolds in Project Databases of Lead Molecules. *J. Med. Chem.* **2009**, *52*, 469–483.

(16) 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.

(17) 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.

(18) Lipkus, A. H.; Yuan, Q.; Lucas, K. A.; Funk, S. A.; Bartelt, W. F.; Schenck, R. J.; Trippe, A. J. Structural Diversity of Organic Chemistry. A Scaffold Analysis of the CAS Registry. *J. Org. Chem.* **2008**, *73*, 4443–4451.

(19) Lameijer, E.-W.; Kok, J. N.; Bäck, T.; IJzerman, A. P. Mining a Chemical Database for Fragment Co-occurrence: Discovery of "Chemical Clichés". *J. Chem. Inf. Model.* **2006**, *46*, 553–562.

(20) Lipkus, A. H. Exploring Chemical Rings in a Simple Topological-Descriptor Space. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 430–438.

(21) Wester, M. J.; Pollock, S. N.; Coutsias, E. A.; Allu, T. K.; Muresan, S.; Oprea, T. I. Scaffold Topologies. 2. Analysis of Chemical Databases. *J. Chem. Inf. Model.* **2008**, *48*, 1311–1324.

(22) Ertl, P.; Jelfs, S.; Mühlbacher, J.; Schuffenhauer, A.; Selzer, P. Quest for the Rings: In Silico Exploration of Ring Universe To Identify Novel Bioactive Heteroaromatic Scaffolds. *J. Med. Chem.* **2006**, *49*, 4568–4573.

(23) Krier, M.; Bret, G.; Rognan, D. Assessing the Scaffold Diversity of Screening Libraries. *J. Chem. Inf. Model.* **2006**, *46*, 512–524.

(24) Hu, Y.; Wassermann, A. M.; Lounkine, E.; Bajorath, J. Systematic Analysis of Public Domain Compound Potency Data Identifies Selective Molecular Scaffolds Across Druggable Target Families. *J. Med. Chem.* **2010**, *53*, 752–758.

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

(26) Medina-Franco, J. L.; Martinez-Mayorga, K.; Bender, A.; Scior, T. Scaffold Diversity Analysis of Compounds Data Sets Using an Entropy-Based Measure. *QSAR Comb. Sci.* **2009**, *11*−*12*, 1551–1560.

(27) Pitt, W. R.; Parry, D. M.; Perry, B. G.; Groom, C. R. Heteroaromatic Rings of the Future. *J. Med. Chem.* **2009**, *52*, 2952–2963.

(28) Clarke, A. 2D Depiction of Fragment Hierarchies. *J. Chem. Inf. Model.* **2010**, *50*, 37–46.

(29) Ghose, A. K.; Crippen, G. M. Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships I. Partition Coefficients as a Measure of Hydrophobicity. *J. Comput. Chem.* **1986**, *7*, 565–577.

(30) EBI-ChEMBL. https://www.ebi.ac.uk/chembldb/index.php/group (accessed March 25, 2011).
(31) ChEMBL_03, ChEMBL-EBI. http://www.ebi.ac.uk/chembldb/index.php (accessed May 14, 2010).
(32) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2008**, *36*, D901–D906.
(33) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastav, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006**, *34*, D668–D672.
(34) DrugBank. http://www.drugbank.ca/downloads (accessed May 14, 2010).
(35) BioFocus. http://www.biofocus.com (accessed March 25, 2011).
(36) *Pipeline Pilot* 7.0; Accelrys: San Diego, CA, USA. http://accelrys.com/ (accessed March 25, 2011).
(37) *MOE* 2009.10; Chemical Computing Group: Montreal, Quebec, Canada. http://www.chemcomp.com/ (accessed March 25, 2011).
(38) Shneiderman, B. Tree visualization with tree-maps: 2-d space-filling approach. *ACM T. Graphic* **1992**, *11*, 92–99.
(39) Kibbey, C.; Calvet, A. Molecular Property eXplorer: A Novel Approach to Visualizing SAR Using Tree-Maps and Heatmaps. *J. Chem. Inf. Model.* **2005**, *45*, 523–532.
(40) *TreeMap* v. 1.9.21; Macrofocus. http://www.macrofocus.com/public/products/treemap/ (accessed March 25, 2011).
(41) Kelley, L. A.; Gardner, S. P.; Sutcliffe, M. J. An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies. *Protein Eng.* **1996**, *9*, 1063–1065.
(42) Schuffenhauer, A.; Brown, N.; Ertl, P.; Jenkins, J. L.; Selzer, P.; Hamon, J. Clustering and Rule-Based Classifications of Chemical Structures Evaluated in the Biological Activity Space. *J. Chem. Inf. Model.* **2007**, *47*, 325–336.