Journal of Chemical Information and Modeling

Article

pubs.acs.org/jcim

Intuitive Ordering of Scaffolds and Scaffold Similarity Searching Using Scaffold Keys

Peter Ertl*

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

ABSTRACT: Scaffold Keys—scaffold descriptors based on simple topological parameters such as number of ring and chain atoms, number and type of heteroatoms, and other simple structural features—are presented. Scaffold Keys enable intuitive ordering of scaffolds from small and simple to large and complex, ordering that is consistent with the way medicinal chemists themselves classify scaffolds. Scaffold Keys may be also used as descriptors for scaffold similarity searches, providing results compatible with expectations of chemists and well-suited for use in scaffold bioisosteric replacement and scaffold hopping. Scaffold Keys also support visualization of large chemical data sets. Scaffold Keys descriptors are easy to understand by chemists as well as easy to implement.

INTRODUCTION

One of the most common tasks that cheminformatics experts in the pharmaceutical or agrochemical industry are facing practically daily is the analysis and visualization of large collections of molecules. Typical areas where this is needed include analysis of the company compound archive and its enhancement by purchasing additional molecules from commercial compound providers, analysis of results of high-throughput screening, design of combinatorial libraries, chemogenomics analysis of bioactivity data, and many others.[1,2] Researchers in academia are facing similar challenges when processing and visualizing large open molecular databases that have become available recently[3−5] or were even generated in silico.[6,7] Ideally, the results of such an analysis should be offered in a visually friendly way, such as a 2D map or graph, taking advantage of the very good capability of the human brain to process graphical information and identify complex patterns in images. Unlike other scientific disciplines, where objects are visualized simply as points on a graph, chemistry visualization, however, presents additional challenges. Chemists want to see also structures of their molecules. This particular requirement makes chemistry visualizations challenging because of the necessity to squeeze a lot of information on rather limited computer screen real estate.

Even for relatively small data sets with a few hundred molecules it is not possible to display all of them on a graph, not to mention large databases, for example results of high throughput screening campaigns where routinely over one million molecules are screened, or the whole company compound archives comprising several million structures. To visualize such large data sets, molecules need to be clustered into meaningful groups, and from each group only a representative example for display selected. A standard way to group molecules is clustering. But, the use of clustering is complicated by the necessity to choose a proper clustering method and respective parameters, both of these choices considerably influencing the results. Clusters generated by many clustering algorithms are not easy to interpret, as every cheminformatics scientist who ever tried to explain to chemists reasons why a molecule belongs to this and not that cluster knows very well. On the other hand, at Novartis we have very good experience with the use of scaffolds as classifying elements. Molecules having the same scaffold are simply placed into one group that is then represented by the respective scaffold. Results of such grouping are easy to understand and are well accepted by chemists, because chemists also tend to think in terms of scaffolds. Scaffolds indeed form a common language between synthetic and medicinal chemists on one side and cheminformaticians and molecular modellers on the other side. Classification of molecules by their scaffolds offers advantages also from the computational point of view. While the run time of most clustering algorithms depends in a quadratic way on the size of the data set and the whole clustering procedure

Received: March 31, 2014

Table 1. Thirty-Two Scaffold Keys

n	description	scope[a]	average[b]	stdev[b]
1	number of ring and linker atoms	RL	20.029	7.556
2	number of linker atoms	L	2.518	3.481
3	number of linker bonds	L	3.993	3.897
4	number of rings		3.348	3.156
5	number of spiro atoms	R	0.031	0.193
6	size of the largest ring	R	6.241	1.905
7	number of bonds in fully conjugated rings	R	13.824	6.201
8	number of multiple bonds in not fully conjugated rings	R	0.112	0.383
9	number of heteroatoms in rings	R	2.177	1.640
10	number of heteroatoms other than N, S, O in rings	R	0.003	0.061
11	number of S ring atoms	R	0.143	0.388
12	number of O ring atoms	R	0.310	0.703
13	number of N ring atoms	R	1.721	1.507
14	number of heteroatoms	A	4.248	2.921
15	number of heteroatoms other than N, S, O	A	0.009	0.131
16	number of S atoms	A	0.289	0.540
17	number of O atoms	A	1.603	1.695
18	number of N atoms	A	2.347	1.789
19	number of multiple linker bonds	A	0.109	0.351
20	count of two adjacent heteroatoms	AO	0.575	1.162
21	count of 3 adjacent heteroatoms	AO	0.350	1.169
22	count of 2 heteroatoms separated by a single carbon	AO	1.804	1.953
23	count of 2 heteroatoms separated by 2 carbons	AO	1.505	2.564
24	number of double bonds with at least one heteroatom	AO	1.235	1.433
25	number of heteroatoms adjacent to a double (nonaromatic) bond	AO	1.439	1.861
26	count of pairs of conjugated double (nonaromatic) bonds	AO	0.094	0.380
27	count of pairs of adjacent branched atoms	AO	2.860	2.320
28	count of branched atoms separated by a single nonbranched atom	AO	1.504	1.467
29	count of 3 adjacent branched atoms	AO	1.734	2.591
30	count of branched atoms separated by any 2 atoms	AO	4.294	4.409
31	number of exocyclic and exolinker atoms	!R, !L	1.170	1.425
32	number of heteroatoms with more than 2 bonds	R	0.673	0.840

[a] R—ring atoms excluding exocyclic atoms; L—linker atoms, excluding exolinker atoms; A—all atoms; O—specified substructures may overlap. [b] Average and standard deviation were calculated for the 10 000 most frequent scaffolds from the ChEMBL database.

needs to be rerun when new molecules are added, scaffold clustering depends in a linear way on the data set size and updates may be handled incrementally. Several methods developed at Novartis based on analysis of scaffolds, including Scaffold Tree,[8] Scaffold Networks,[9] or visualization of large data sets using Molecule Clouds[10] document advantages of this approach.

The method based on scaffolds, however, also has its disadvantages. Reducing molecules only to their scaffolds neglects the substituent pattern that is in many cases crucial for biological activity. However, if a specific scaffold keeps the pharmacophoric side chains in the right geometric arrangement, then this scaffold represents a common core in the whole class of active compounds, although it is not the pharmacophore itself. In any case one should keep in mind that the main purpose of the presented method is visualization of large chemical data sets and not overinterpret implications on biological activity.

Both Scaffold Tree and Molecule Cloud diagrams provide very useful visualization of chemical space; they have, however, a common disadvantage. Although these methods offer clear and in most cases also aesthetically pleasing visualization of most important scaffolds in a data set, and in cases of Scaffold Tree also hierarchical relationships between scaffolds, the axes of display area have no physical meaning; scaffolds are ordered in the graphs more or less randomly. It would be of advantage, if the display would respect also some intuitive ordering of scaffolds. Indeed, more than 20 years of the author's experience in supporting numerous agrochemical and pharmaceutical projects clearly shows that medicinal chemists automatically classify molecules when they see them, mostly based on the scaffolds. Every chemist uses, of course, slightly different "classification rules" based on his/her training and experience, and it would be difficult to exactly formulate these intuitive rules, but the basic principles of such classification are the same. The most obvious discriminant feature is the size of the scaffold. Other commonly used characteristics are its aromatic or aliphatic character and number and types of heteroatoms. Also the presence of noncommon structural features, like spiro centers or nonclassical fusing, when two rings share more than two common atoms, plays a role.

Several attempts to use simple structural features to classify molecules and scaffolds have been reported in the scientific literature. Nilakantan[11] suggested a simple scaffold complexity score for scaffolds classification calculated only from the number of atoms and bonds in the system. The score is easy to calculate, but sometimes very different scaffolds are grouped together due to frequent collisions in the score. Lipkus presented a method for classification of chemical rings.[12] The method uses three simple descriptors derived from the ring topologies and allows placing any ring to a cell in a 3D "ring topology" space. Analysis of the Chemical Abstract Registry data provided over 40 000 different ring topologies. This method, however, did not take into account heteroatoms or different bond orders in the processed rings. Katritzky[13] suggested a sophisticated method to quantify differences between scaffolds with the goal to classify combinatorial libraries. He introduced several simple rules taking into account differences like subtraction or replacement of atoms, changes to rings fusion and introduction of spiro rings. Very popular descriptors for classification of molecules are MDL Keys,[14] bits representing presence or absence of simple structural features like atoms or bonds of certain type or various substructures. Nguyen et al. introduced so-called molecule quantum numbers to classify molecules.[15] These numbers are actually counts of 42 simple structural features like atoms or bonds of certain type or polarity counts. Molecule quantum numbers are able to cluster compounds with similar structure, physicochemical properties, and bioactivities. Also an interactive molecule browser based on principal component analysis of molecule quantum numbers was presented.[16] Li et al. introduced a sophisticated method to determine the structural distance between scaffolds.[17] Scaffolds to be compared are transformed to each other by a series of editing steps until equivalence is achieved. To every step, an empirical score to quantify the change is assigned. The method allows the evaluation of the distance between scaffolds in scaffold hopping applications. Five simple topological keys to classify over 100 000 possible "non-terrestrial amino acids" were also recently described.[18]

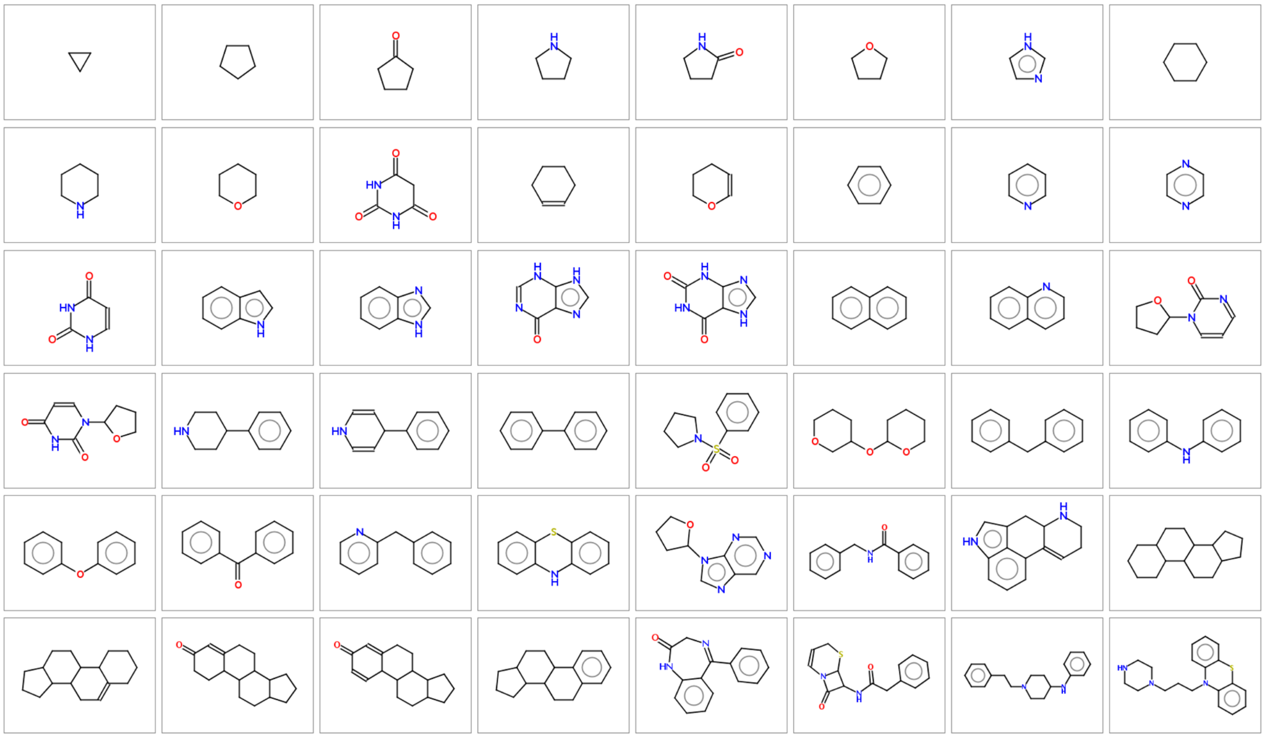

Figure 1. Most common scaffolds from the ChEMBL database ordered according to their Scaffold Keys.

METHODOLOGY

The method for classification of scaffolds presented here is based on simple topological descriptors: counts of atoms and bonds of particular types and various substructure scaffold features characterizing ring fusing, spiro connections, and linker properties. Before providing details about the procedure, however, let us focus in more detail on the definition of the term "scaffold". This word is used very often in medicinal chemistry literature. Despite its importance, however, it is used rather freely, without a clearly defined meaning. The exact interpretation of this term varies from publication to publication and depends also on the particular application area. In some cases, the term scaffold is used for a part of molecule remaining after removal of all nonring substituents (such scaffolds are often called also Bemis−Murcko scaffolds[19] or molecule frameworks). Medicinal chemists usually understand by scaffold, the largest central ring system in the molecule, while smaller rings at the periphery (often called "decorations") form a SAR pattern. When used in the combinatorial chemistry applications the term scaffold is used to describe a substructure common to all molecules in a library and may contain rings as well as nonring functionalities, in extreme cases even lacking rings at all. This nomenclature ambiguity makes therefore necessary in every application a clear definition of what is understood by the term "scaffold". Throughout this article the term scaffold is used to describe part of the molecule that remains after removal of nonring substituents, keeping, however also exocyclic and exochain multiple bonds.

As mentioned previously, in the method presented here scaffolds are characterized by a set of simple and easy to understand topological descriptors that we term Scaffold Keys. The list of these descriptors together with some additional data are listed in Table 1. The keys are ordered according to their perceived importance by chemists; the keys characterizing size of scaffolds and their ring composition are the most important, then come keys characterizing presence and types of heteroatoms, and finally the keys that can distinguish scaffold isomers.

The actual descriptors that form the Scaffold Keys set were selected empirically after a long trial and error procedure and discussions with chemists to provide the best description of various scaffold features, optimal grouping of scaffolds, and minimal number of collisions.

The actual implementation of Scaffold Keys was done by an in-house code written in Java. The code itself, since it is part of a larger internal code base, unfortunately cannot be shared, but the reimplementation of Scaffold Keys by using any standard cheminformatics toolkit should not be difficult. The author is willing to provide advice to any scientist trying to reimplement the code.

INTUITIVE ORDERING OF SCAFFOLDS

The primary application of Scaffold Keys is ordering of scaffolds. Scaffolds are sorted first according to the first key, then the second key, and so on to obtain what may be called an "intuitive scaffold ordering". Examples of such ordering for the 48 most frequent scaffolds present in the ChEMBL database[4] are shown in Figure 1. One can see that scaffolds are indeed ordered from

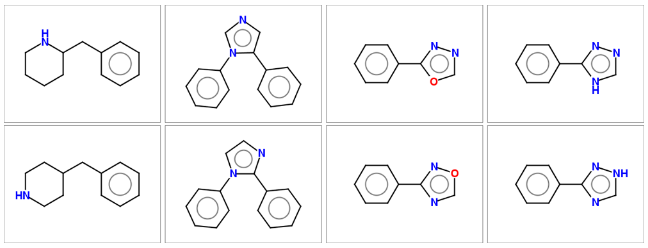

Figure 2. Examples of four pairs of scaffolds having the identical Scaffold Keys. The scaffolds are indeed very similar.

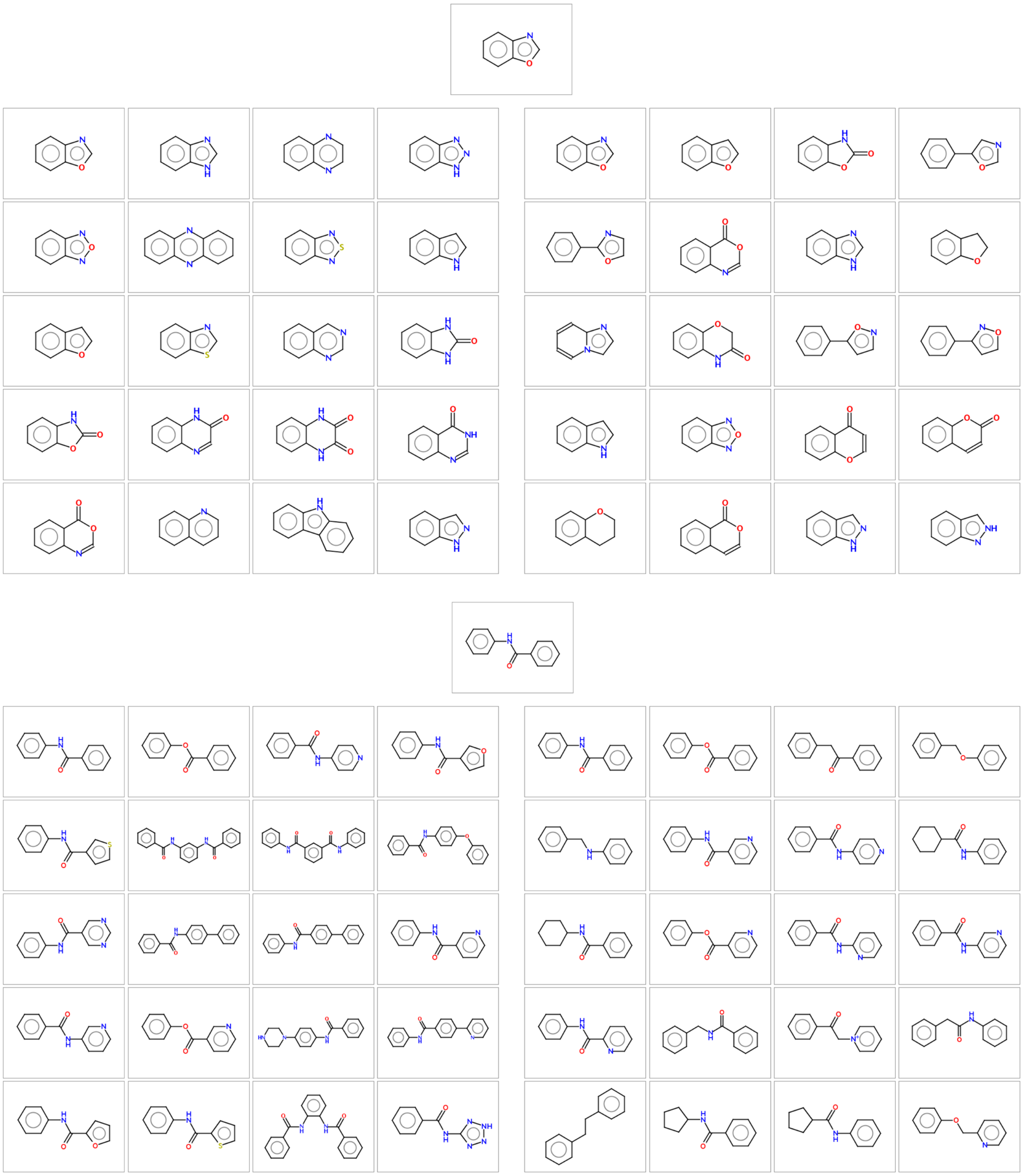

Figure 3. Comparison of results of classical similarity search (left blocks) with Scaffold Keys similarity (right blocks) for two queries (see the main text for details).

small and simple to large and complex, ordering that is consistent with the way the medicinal chemists themselves intuitively classify scaffolds.

Of course, when using only 32 simple integer keys to characterize scaffolds, one has to accept that collisions will happen, i.e., different scaffolds will have the same Scaffold Keys. Examples of different scaffolds having the same Scaffold Keys are shown in Figure 2, but one can see that these scaffolds are indeed very similar to each other, being either tautomers or isomers differing in position of heteroatoms.

We are aware of the fact that different cheminformatics toolkits may provide slightly different values of some keys, depending on the internal molecule representation used by various toolkits. This is particularly true for keys related to identification of aromatic bonds (keys 25 and 26), since the exact definition of "aromaticity" (and its software implementation in various toolkits) is still a topic not generally agreed on. But since such keys are rather low in precedence, slight differences in their values will influence results of scaffold ordering and similarity

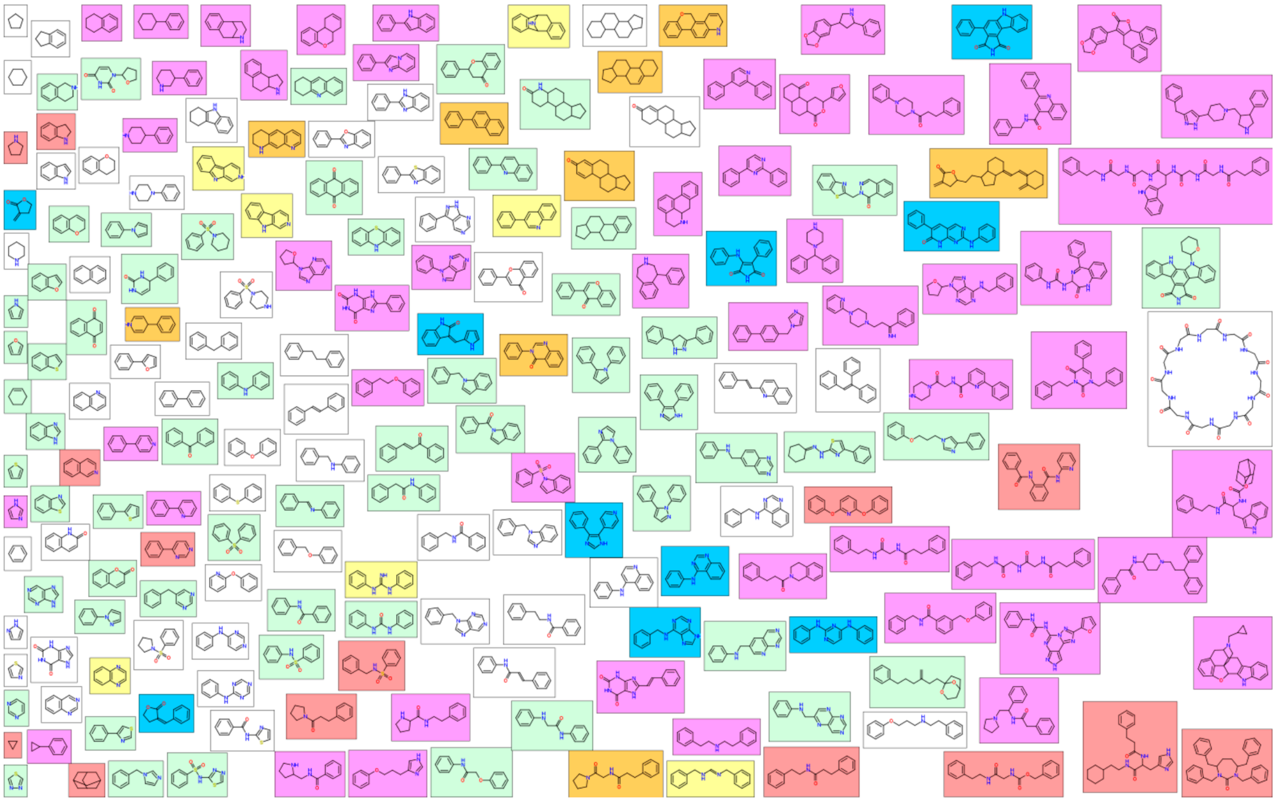

Figure 4. Scaffold Map for the most frequent scaffolds in the ChEMBL database. The scaffold preference for a particular target class is indicated by color (see main text for details).

searches only marginally and the keys will separate closely related scaffolds and minimize collisions of the whole key set, anyway.

SCAFFOLD SIMILARITY SEARCHES

One of the major motivations to develop a specific set of scaffold descriptors was to use it in scaffold similarity searches. Classical molecular similarity searches based on fingerprints and Tanimoto similarity measure that are so successful for standard drug-like molecules[20] are unfortunately not working well for scaffolds. This is caused by the fact that scaffolds are relatively small chemical objects with limited number of distinct structural features that set only few bits in the respective fingerprints. The Tanimoto coefficient used for actual calculation of similarity score considers only presence or absence of structural features and not their count. Therefore, the top hits provided by such "classical" similarity searches contain indeed the same structural features as the query; however, these are often present multiple times, and hits may be then considerably larger than query scaffolds. This is illustrated by comparing results of classical similarity search (using Schrödinger Canvas[21] with radial fingerprints, default atom typing, and Tanimoto similarity coefficient) with similarity search based on Scaffold Keys for two simple queries as shown in Figure 3. Also additional "classical" similarity search settings and also tools of other software vendors were tried, but results were quite similar to those shown on the left side of Figure 3. Perception of similarity is, of course, very subjective and in some cases a certain level of diversity in search results may be even beneficial. But one can clearly see that Scaffold Keys similarity respects size and shape scaffold properties, i.e. properties that play the most important role in scaffold bioisosteric replacement, better than the "classical" similarity search.

The calculation of scaffold similarity (actually a distance between scaffold *i* and scaffold *j*) based on Scaffold Keys is described by the following pseudocode:

```
# distance between scaffold i and scaffold j
distance = 0.
for n = 1; n <= 32; n++: # loop through 32 Scaffold Keys
  # contributions of keys are weighted by the key number
  distance += (abs(key[i][n] - key[j][n])**1.5) / n
```

To calculate the distance the original Scaffold Keys integer values need to be normalized to have an average of 0 and standard deviation of 1. For the normalization statistical parameters generated for the 10 000 most frequent scaffolds from the ChEMBL database were used. These parameters are also included in Table 1. The output of the formula is zero when the two scaffolds are identical (have the same Scaffold Keys) and increases with the difference in keys. Scaffolds identified by similarity searches based on Scaffold Keys may be used with advantage in scaffold hopping applications or scaffold bioisosteric replacement.[22]

VISUALIZATION OF LARGE CHEMICAL DATA SETS USING SCAFFOLD MAPS

The Scaffold Keys descriptors are useful also to visualize large collections of molecules. As mentioned already in the introduction, if a large number of molecules need to be visualized, the molecules should be grouped together based on the common scaffold and this scaffold then represents the whole group. The scaffolds may be arranged on a computer screen (or on a printed graph) using their Scaffold Keys. The horizontal axis of the graph represents the first key, on the vertical axis scaffolds are displayed based on the ordering according to all remaining keys. The horizontal axis represents therefore scaffold size, the vertical axis is a mixture of other scaffold characteristics and may be therefore roughly described as a measure of complexity or feature richness. In displaying the scaffolds on a limited plot area, one has to find a compromise between the exact position of scaffolds as determined by their Scaffold Keys and the best use of available space. After placing scaffolds in original positions in the graph therefore an layout optimization procedure described in ref 10 is used, i.e., scaffolds are slightly iteratively moved to minimize their overlap. The example of the resulting graph that we call Scaffold Map, is shown in Figure 4. One can also use color to show additional information. In this case, the color represents preferred target classes for molecules containing this scaffold. The preference of a scaffold for a particular target class is indicated by color (magenta—GPCRs; blue–kinases; red–proteases; green–other enzymes; brown–nuclear receptors; yellow–ion channels), noncolored scaffolds show activity on multiple targets (data from the ChEMBL database[4]). According to our experience it is possible to display in this way 150–300 scaffolds on a standard PC computer screen. The display of Scaffold Map may be enhanced by adding additional interactivity, for example clicking on the scaffold depiction on a computer screen can open an additional window with information about all molecules containing this scaffold.

■ CONCLUSIONS

The Scaffold Keys—a set of simple topological descriptors developed specially to characterize scaffolds—is described. Scaffold similarity searches based on Scaffold Keys provide results consistent with expectations of medicinal chemists, respect the size and shape of scaffolds, and, therefore, are well-suited for scaffold hopping applications and scaffold bioisosteric searches. Scaffold Keys may be used also to sort scaffolds and to visualize large molecular data sets by Scaffold Maps. The Scaffold Key descriptor is simple and should be easy to implement by using any standard cheminformatics toolkit.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: peter.ertl@novartis.com. Web address: http://peter-ertl.com

Notes

The author declares no competing financial interest.

■ ACKNOWLEDGMENTS

The author wants to thank Paulette Greenidge for critically reading the manuscript and for helpful comments.

■ REFERENCES

(1) Irwin, J. J. Staring off into Chemical Space. *Nature Chem. Biol.* **2009**, *5*, 536–537.

(2) Medina-Franco, J. L.; Martinez-Mayorga, K.; Giulianotti, M. A.; Houghten, R. A.; Pinilla, C. Visualization of the Chemical Space in Drug Discovery. *Curr. Comp.-Aided Drug Des.* **2008**, *4*, 322–333.

(3) Wang, Y.; Bolton, E.; Dracheva, S.; Karapetyan, K.; Shoemaker, B. A.; Suzek, T. O.; Wang, J.; Xiao, J.; Zhang, J.; Bryant, S. H. An Overview of the PubChem BioAssay Resource. *Nucleic Acids Res.* **2009**, *38*, D255–D266.

(4) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a Large-scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–D1107.

(5) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757–1768.

(6) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875.

(7) Ertl, P.; Jelfs, S.; Muehlbacher, 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.

(8) 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.* **2011**, *51*, 1528–1538.

(9) Varin, T.; Schuffenhauer, A.; Ertl, P.; Renner, S. Mining for Bioactive Scaffolds with Scaffold Networks: Improved Compound Set Enrichment from Primary Screening Data. *J. Chem. Inf. Model.* **2007**, *47*, 47–58.

(10) Ertl, P.; Rohde, B. The Molecule Cloud - Compact Visualization of Large Collections of Molecules. *J. Cheminf.* **2012**, *4*, 12.

(11) Nilakantan, R.; Bauman, R.; Haraki, K. S.; Venkataraghavan, R. A Ring-based Chemical Structural Query System: Use of a Novel Ring-complexity Heuristic. *J. Chem. Inf. Comp. Sci.* **1990**, *30*, 65–68.

(12) 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.

(13) Katritzky, A. R.; Kiely, J. S.; Hébert, N. Chassaing, Ch. Definition of Templates Within Combinatorial Libraries. *J. Comb. Chem.* **1999**, *2*, 2–5.

(14) Durant, J. L.; Burton, A. L.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Model.* **2002**, *42*, 1273–1280.

(15) Nguyen, K. T.; Blum, L. C.; van Deursen, R.; Reymond, J.-L. Classification of Organic Molecules by Molecular Quantum Numbers. *ChemMedChem.* **2009**, *4*, 1803–1805.

(16) Ruddigkeit, L.; Blum, L. C.; Reymond, J.-L. Visualization and Virtual Screening of the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2013**, *53*, 56–65.

(17) Li, R.; Stumpfe, D.; Vogt, M.; Geppert, H.; Bajorath, J. Development of a Method To Consistently Quantify the Structural Distance Between Scaffolds and To Assess Scaffold Hopping Potential. *J. Chem. Inf. Model.* **2011**, *51*, 2507–2514.

(18) Meringer, M.; Cleaves, H. J.; Freeland, S. J. Beyond Terrestrial Biology: Charting the Chemical Universe of α-Amino Acid Structures. *J. Chem. Inf. Model.* **2013**, *53*, 2851–2862.

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

(20) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comp. Sci.* **1998**, *38*, 983–996.

(21) *Canvas*, version 1.6.047, Schrödinger, Inc.: New York.

(22) Ertl, P. Database of Bioactive Ring Systems with Calculated Properties and its Use in Bioisosteric Design and Scaffold Hopping. *Bioorg. Med. Chem.* **2012**, *20*, 5436–5442.