The Scaffold Tree − Visualization of the Scaffold Universe by Hierarchical Scaffold Classification

Ansgar Schuffenhauer,*,† Peter Ertl,† Silvio Roggo,† Stefan Wetzel,‡ Marcus A. Koch,‡ and Herbert Waldmann‡

Novartis Institutes for BioMedical Research, CH-4002 Basel, Switzerland, and Max Planck Institute of Molecular Physiology and Fachbereich 3 − Chemical Biology, University of Dortmund, D-44227 Dortmund, Germany

Received August 2, 2006

A hierarchical classification of chemical scaffolds (molecular framework, which is obtained by pruning all terminal side chains) has been introduced. The molecular frameworks form the leaf nodes in the hierarchy trees. By an iterative removal of rings, scaffolds forming the higher levels in the hierarchy tree are obtained. Prioritization rules ensure that less characteristic, peripheral rings are removed first. All scaffolds in the hierarchy tree are well-defined chemical entities making the classification chemically intuitive. The classification is deterministic, data-set-independent, and scales linearly with the number of compounds included in the data set. The application of the classification is demonstrated on two data sets extracted from the PubChem database, namely, pyruvate kinase binders and a collection of pesticides. The examples shown demonstrate that the classification procedure handles robustly synthetic structures and natural products.

INTRODUCTION

The concept of a chemical scaffold as a common core structure characterizing a group of individual molecules in which it is contained as a substructure has a long tradition in chemistry. Structures sharing a scaffold can often be assumed to share a common synthetic pathway; and in typical combinatorial libraries all compounds are based on a common scaffold. Scaffolds are used to define classes of chemical compounds in patent claims which are referred to as Markush structure.[1] Therefore, the modification of a structure in such a way that its scaffold is changed, but its desirable properties like biologic activity are retained, is of high value and often referred to as scaffold hopping.[2,3] In the analysis of biological screening data, it is of interest to group compounds on the basis of common core scaffolds. From the individual compounds' substitution patterns structure−activity relationship (SAR) information may subsequently be derived which guides the further optimization of bioactivity of this scaffold.

To group data sets of chemical structures by their scaffolds requires a definition of chemical scaffolds which allows encoding of a computational procedure to extract the scaffold out of a chemical structure. The molecular framework as a useful definition of the scaffold has been introduced by Bemis and Murcko.[4] It is defined as the part of a structure which remains after all terminal chains have been removed. By discarding from the molecular framework atom- and bond-type information step by step and ultimately condensing the connectivity information to a reduced graph, a series of scaffolds can be obtained with increasing abstraction which can be used for the hierarchical classification of chemical structures. A canonical numbering for framework graphs with different levels of abstraction has been introduced as molecular equivalence numbers (MEQNum).[5] This procedure has been used to identify scaffolds related to biological activity[6] and to define a measure of structural diversity within a data set.[7] While these definitions have been shown to provide useful results in the cases published, the abstraction of atom, bond, and ring size information from molecular frameworks generates artificial, nonmolecular entities instead of substructures of the original molecules. Another drawback of the concept using molecular frameworks cannot be addressed with these procedures: the addition of a cyclic substituent will always change the scaffold of the molecule itself, and this change is preserved in all abstractions of the framework. This problem has been addressed by Katritzky et al.,[8] who have devised rules when scaffolds should be considered as equivalent. Two scaffolds are equivalent if the summed scores of the transformations needed to convert one into the other according to the proposed scoring scheme does not exceed a given threshold. However, because no canonical representation of all equivalent scaffolds has been proposed, this concept cannot be used for automated classification. Another procedure, called HierS,[9] groups the molecular frameworks hierarchically on the basis of the ring systems contained in the scaffolds, which are obtained when all linker bonds are removed. Each individual ring system and each combination of them defines a class to which the scaffold is assigned. This means, however, that a scaffold is assigned to more than one class, making their use for data analysis, especially in the case of larger data sets, more complicated. Also, HierS does not dissect fused ring systems and can therefore not simplify complex fused or bridged cyclic systems often occurring in natural products. Another, even more complex data analysis tool, called Leadscope, uses a manually built, hierarchically sorted dictionary of cyclic and

* Corresponding author e-mail: ansgar.schuffenhauer@novartis.com.
† Novartis Institutes of Biomedical Research.
‡ University of Dortmund.

10.1021/ci600338x CCC: $37.00
Published on Web 12/02/2006

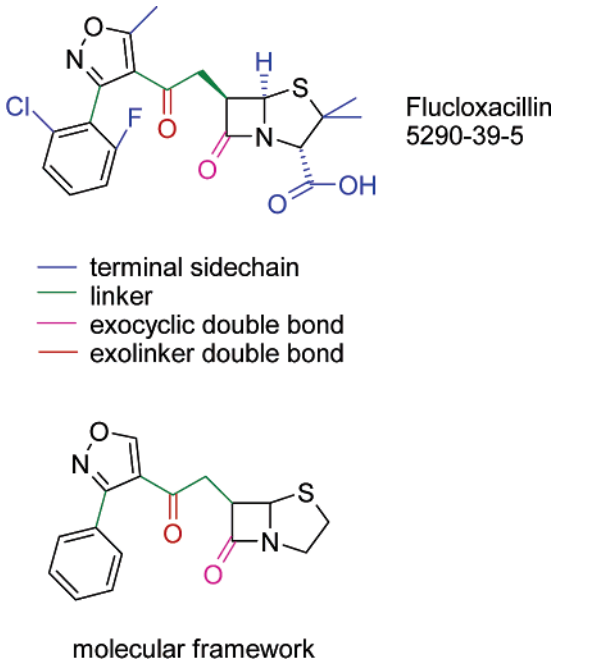
acyclic fragments to analyze structural data sets. Also, this system is based on the assignment of structures to multiple groups.[10]

In this article, a method for the classification of cyclic scaffolds is described which is based on dissection of the scaffolds by an iterative removal of rings, until a single "root" ring is obtained. At each iteration step, prioritization rules are applied to decide which ring to remove. This leads to a unique, hierarchical classification of scaffolds, where each scaffold in the hierarchy is a well-defined chemical entity, which is contained in the original molecule as a substructure. Therefore, established procedures such as canonical SMILES[11] can be used to generate a canonical representation for each subscaffold in the hierarchy tree. The uniqueness however has its price: As each scaffold in the classification tree can have only one parent scaffold, one has to select the prioritization rules carefully in order to retain that part of the scaffold as a parent which characterizes it in a chemically intuitive way. This usually means retaining central and complex rings and removing peripheral simple rings. In general, the goal of the method is to obtain with our rule set a chemically meaningful classification and not a classification with respect to pharmacophoric elements.

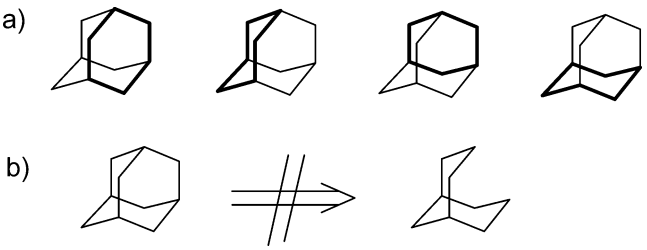
We applied an early version of this method in the analysis of natural product structures[12] and combined it with a structural classification of proteins[13] to introduce the concept of biology oriented synthesis.[14] Because in this predecessor method the frequency of occurrence of parent molecules was one of the criteria used to decide which rings to retain, this provided some in-built protection preventing the selection of chemically nonintuitive subscaffolds for the classification. The disadvantage, however, was that the classification method was not data-set-independent, meaning that the further addition of structures to the data set could change the outcome of the classification of the whole set. Therefore, in further development of this approach, we fine-tuned the prioritization rules in order to be able to dismiss the frequency as a prioritization criterion.

This refined rule set is described here. We illustrate the method by applying it to two sets of bioactive molecules obtained from the PubChem database.

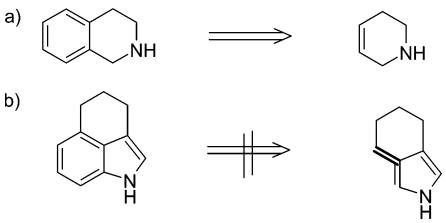
Scheme 1



Scheme 2



Scheme 3



SCAFFOLD CLASSIFICATION PROCEDURE

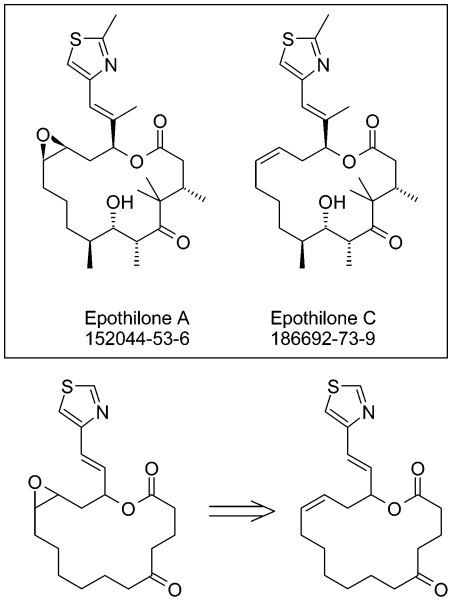
The classification procedure, which has been implemented in-house in Java by using the Molinspiration toolkit (www.molinspiration.com) begins by removing all terminal side chains to obtain the molecular framework. Exocyclic double bonds, and double bonds directly attached to the linker ("exolinker double bonds") are kept (Scheme 1). This is to ensure that planar sp^2 carbon atoms are recognizable in the scaffold and are not converted into tetrahedral sp^3 carbon atoms and thereby changing the local geometry of the ring or linker. The stereochemistry is discarded at the stage the molecular framework is determined. In an ideal world, one would have retained the information about the configuration of stereocenters as long as possible, but in a real-world scenario, where for many databases stereo information is either not available or incomplete, this would be a source of error as the outcome of the classification would depend on the presence of stereo information. Given the fact that in diversity selection applications 2D descriptors performed better than or equally as well as 3D descriptors, the loss of stereochemistry information can be expected to have little impact on the sampling of the scaffold space as well.[15]

From this scaffold, rings are removed iteratively one by one until only one ring remains. Removal of a ring means that bonds and atoms which are part of the ring are removed excluding atoms and bonds which are part of any other ring. In addition, all exocyclic double bonds attached to the removed ring atoms are removed as well. If the removed ring was connected to the remaining scaffold by an acyclic linker, this linker is now a terminal side chain and is removed as well.

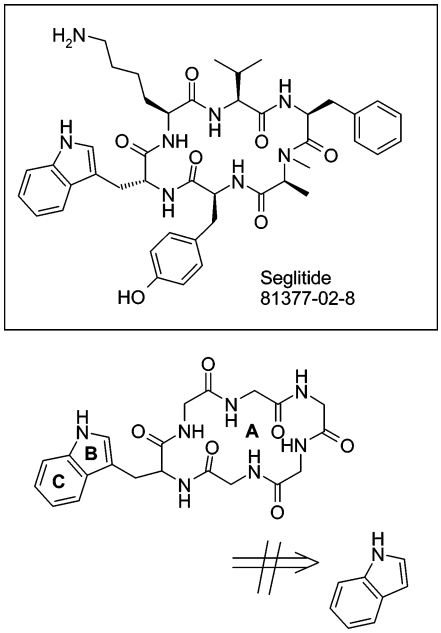
If the removal of a ring would lead to a disconnected structure, this ring cannot be removed.

The outcome of the classification depends on the exact definition of rings and ring membership. For this, we used the ring perception implemented in the Molinspiration toolkit based on the $\mathscr{K}$-ring set, which is obtained as a logical union of all possible smallest sets of smallest rings for the

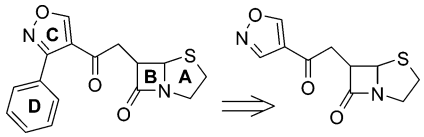
Scheme 4



Scheme 5



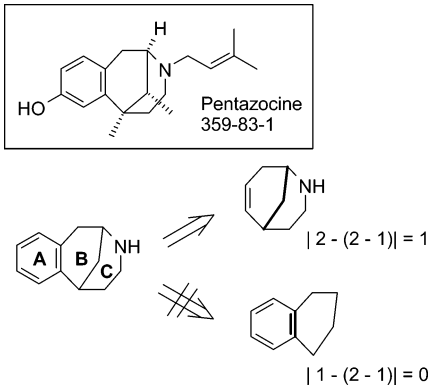
Scheme 6



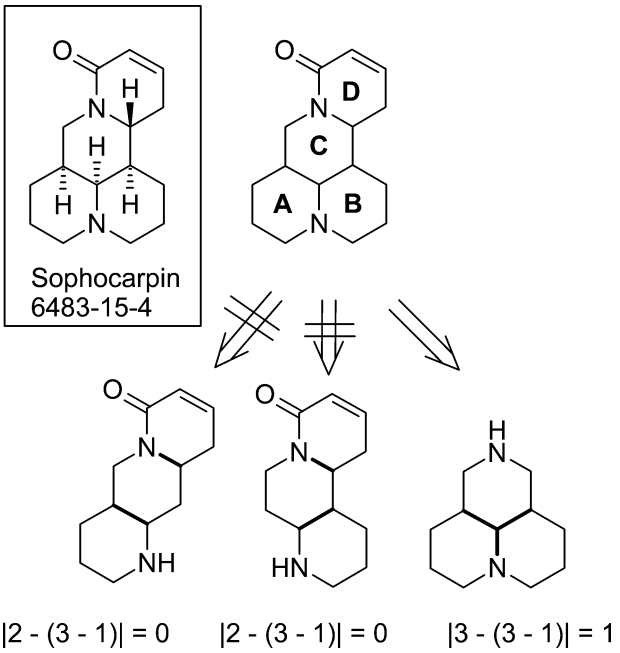
respective molecule.[16] In this set of $\mathscr{K}$-rings, the symmetry of highly symmetric ring systems like adamantane is retained (the $\mathscr{K}$-ring set contains in this case four six-membered rings, unlike three rings selected randomly as in the case of the smallest set of smallest rings; Scheme 2a). In such rare cases where highly symmetric rings are present it is often not possible to remove only one ring at a time, because there are no atoms belonging only to one ring. In this situation the algorithm stops at this stage and the ring system is not dissected further (Scheme 2b).

In case nonaromatic and aromatic rings are fused, and according to the prioritization rules, the aromatic ring is to

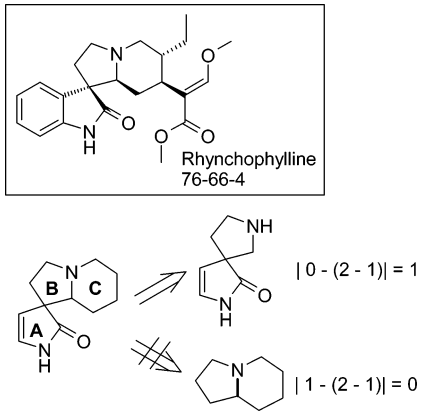
Scheme 7



Scheme 8

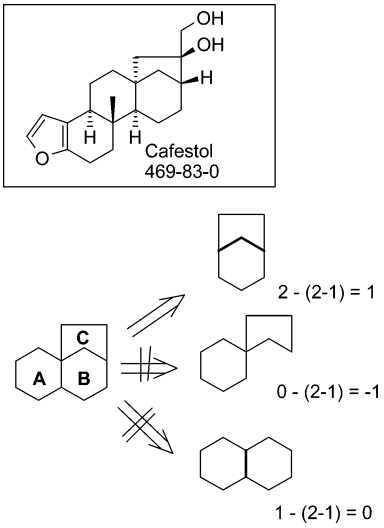


Scheme 9

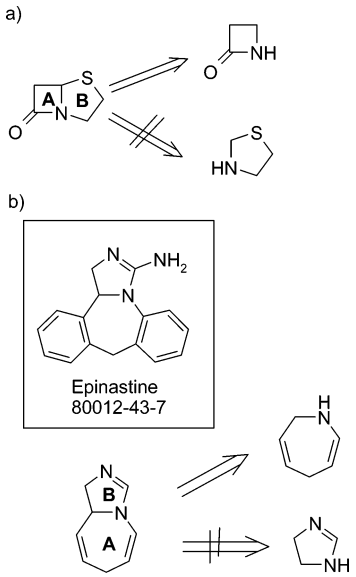


be removed; the remaining, isolated aromatic bond is converted to a double bond. This avoids creation of chemically not meaningful rings and retains the planar geometry of the carbon atoms at the bond where the rings were fused (Scheme 3a). Sometimes, the removal of an aromatic ring leads to an undefined state, such as that shown in Scheme 3b. In this case, the removal of an aromatic rings leads to a remaining aromatic bond which cannot be converted into a

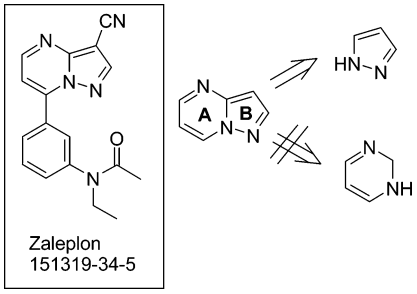
Scheme 10



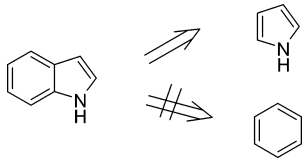
Scheme 11



Scheme 12

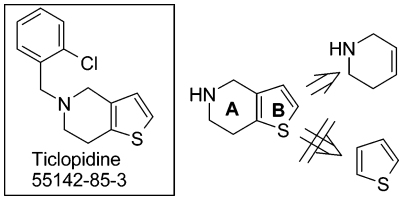


Scheme 13

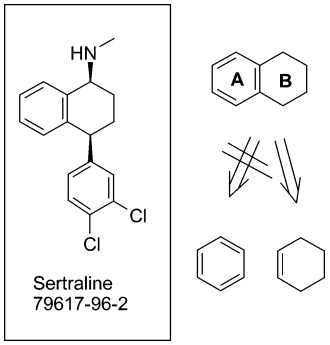


double bond without violating the valence rules. At the same time, leaving the aromatic bond in place would be chemically meaningless as well. If the removal of a ring would lead to such a situation, the ring cannot be removed.

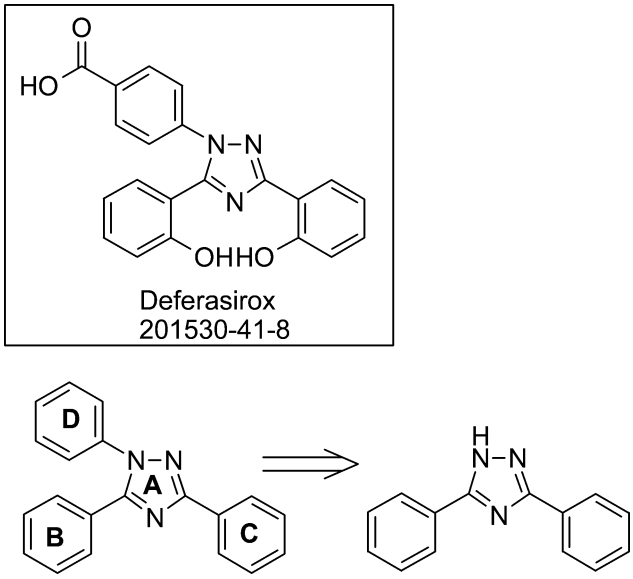
Scheme 14



Scheme 15



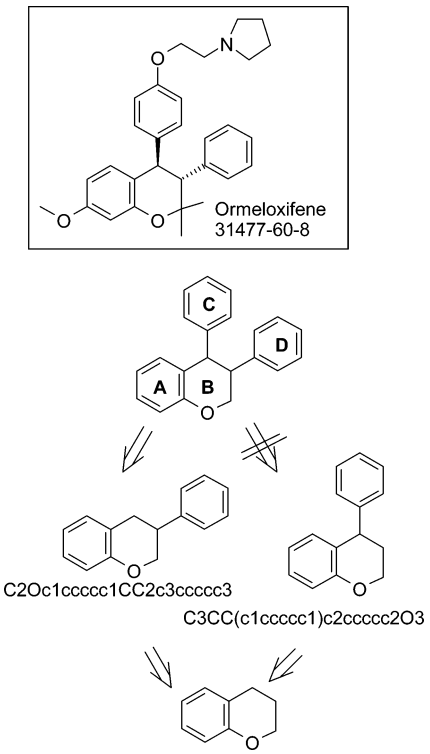
Scheme 16



At each iteration step, the prioritization rules described below are applied to determine which of the removable rings is to be actually removed at this step. The prioritization rules described are listed in the sequence of precedence and are checked in this sequence at each iteration step. As soon as a ring is unambiguously identified for removal, it is removed without checking further rules with lower precedence. Because of the order of precedence, the exceptions are listed before the more generic rules.

1. Remove Heterocycles of Size 3 First. As an exception to the general rules, the fusion bond connecting the three-membered ring with other rings is converted into a double bond. This rule is intended to deal with epoxides and aziridines. This rule treats such systems as functional groups which are removed beforehand, rather than as rings. This reflects the situation that epoxides are usually generated by the oxidation of a double bond, and also many natural products exist often in forms with and without epoxidized double bonds (Scheme 4).

Scheme 17



2. Do Not Remove Rings with ≥ 12 Atoms if There Are Still Smaller Rings To Remove. If a structure contains a macrocycle, this is considered to be the most characteristic ring system occurring in the molecule. Therefore, it should be retained. Especially, cyclic peptides may have bicyclic indole side chains from tryptophane which would be favored by the more general rules below. In the example shown in Scheme 5, either ring A or C can be removed. The removal of ring B would dissect the molecule into disconnected fragments, which is forbidden. According to this rule, we have to preserve ring A, and thus ring C is removed. While macrocyles are retained at the level of the individual macro−monocycles, they are not classified further by the rule set described here. Special rules for the classification of macrocyles have been defined elsewhere.[17]

3. Choose the Parent Scaffold Having the Smallest Number of Acyclic Linker Bonds. This leads to the removal of linked rings before removing fused rings. Rings linked by longer chains are removed first. Linkers are usually the most likely point of a retrosynthetic disconnection. In the synthesis of combinatorial libraries, the variable side chains are often attached to a cyclic core by some linking reaction creating an acyclic linker. Whenever different cyclic side chains are used, their pruning at an early stage leads to the preservation of the common core of the library. Therefore, it is intuitive to dissect scaffolds at acyclic linkers. Also, this helps in retaining preferentially more rigid scaffolds which are more likely to have a unique interaction pattern. In the scaffold of flucloxacillin shown in Scheme 6, only rings A and D could be removed without disconnecting the scaffold. After the removal of ring A, the scaffold would have four acyclic linker bonds; after the removal of ring D, the scaffold has three acyclic linker bonds, and thus this ring is removed.

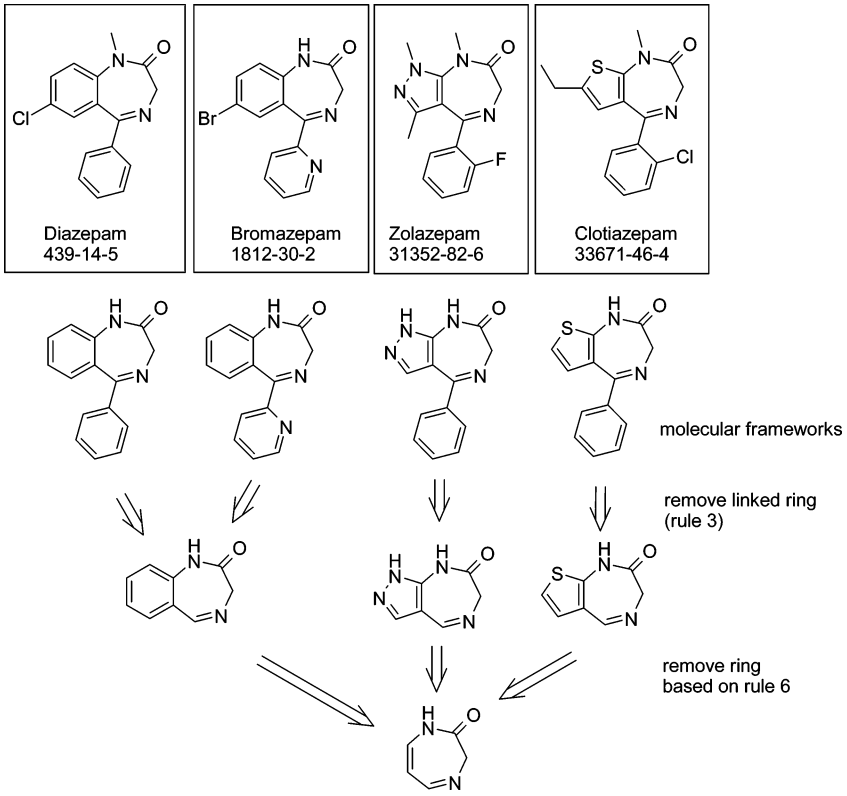
4. Retain Bridged Rings, Spiro Rings, and Nonlinear Ring Fusion Patterns with Preference. These patterns are unusual structural features occurring less frequently than normally fused rings. They have nonplanar, characteristic molecular shapes, which distinguishes them from the majority of the more planar organic molecules. In most ring systems, we have a linear fusion with no atoms in common to more than two rings. This is, for example, the case in steroids. In such cases, the number of bonds being a member in more than one ring n_{rrb} is equal to the number of rings $n_R - 1$. The more bridges or nonlinear ring fusions there are, the higher the number of n_{rrb} is. On the other hand, n_{rrb} decreases if there are spiro connected ring systems, because the spiro connections lead to no bond in common to both rings. Therefore, we remove that ring with preference where the remaining scaffold has the highest value for $|\Delta| = |n_{rrb} - (n_R - 1)|$.

The effect of this rather complex rule is illustrated in three examples. In Scheme 7, the scaffold of pentazocine is shown. Ring A or C can be removed without disconnecting the scaffold. If the bridged BC ring system is retained, there are two bonds being part of more than one ring, leading to $|\Delta| = 1$, whereas the fused bicycle AB leads to $|\Delta| = 0$. Therefore, the BC ring system is retained. In Scheme 8, the scaffold of sophocarpin is shown. Ring A, B, or D can be removed without disconnecting the scaffold. The highest $|\Delta|$ of 1 is obtained when ring D is removed, and the nonlinearly fused tricycle ABC is retained. As a third example, the dissection scaffold of rhynchophylline is shown in Scheme 9. After the removal of the benzene ring in a first dissection step, there remains the ring system ABC of which A or C could be removed. Retaining the spiro-ring system AB leads to $|\Delta| = 1$, whereas retaining the bicycle BC leads to $|\Delta| = 0$. Consequently ring C is removed.

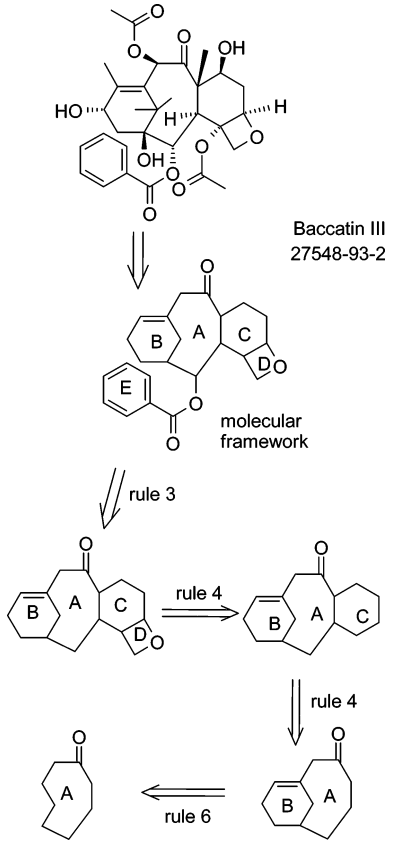
5. Bridged Ring Systems Are Retained with Preference over Spiro Ring Systems. Under certain circumstances, ring systems containing ring fusions as well as bridged rings can be dissected to produce a spiro ring or alternatively a bridged ring. Both solutions would have the same $|\Delta|$ value. A typical example for this situation is shown in Scheme 10. In this situation, it seems to be less intuitive to artificially create a spiro ring system than retaining a bridged ring. Therefore, in the cases where the two remaining subscaffolds have the same value for $|\Delta|$, the ring system with a positive signed value of Δ is to be retained. From the scaffold of cafestol after the removal of two rings, which are identified according to rule 4, the tricyclus ABC is obtained. Already from rule 4 it is clear that ring C must not be removed. However, the bridged ring system BC and the spiro ring system AC have an equal $|\Delta|$ of 1. However, the signed value of Δ is positive for ring system BC and negative for ring system AC. Therefore, ring A is removed.

6. Remove Rings of Sizes 3, 5, and 6 First. Rings of sizes 3, 5, and 6 are more frequent and also synthetically more easily accessible than rings of other sizes. The majority of the commercially available building blocks contain rings of size 3, 5, or 6. If rings of different sizes occur, they are likely to be built up intentionally to fulfill a dedicated purpose. Often, such rings are retained throughout a whole series of bioactive compounds. Good examples for this are penicillin, diazepam, and imipramin together with their related "me-too" analogue compounds. In the bicyclic penam scaffold obtained in the dissection of the scaffold of flucloxacillin, the β-lactam ring A is retained and the five-

Scheme 18



Scheme 19



membered ring B is removed (Scheme 11a). In the same way, in the scaffold remaining after pruning the initial benzene rings from the molecular framework of epinastine, the seven-membered ring A is retained and the five-membered ring B is removed (Scheme 11b).

7. A Fully Aromatic Ring System Must Not Be Dissected in a Way That the Resulting System Is Not Aromatic Any More. The conversion of aromatic in nonaromatic rings is chemically nonintuitive, and it would also affect the geometry of the ring atoms. For example, in the case of zaleplon (Scheme 12) after the initial pruning of the linked phenyl ring, the fully aromatic bicyclic system AB is obtained. From this, the removal of ring B would give a nonaromatic ring, and therefore ring A is removed.

8. Remove Rings with the Least Number of Heteroatoms First. Exocyclic double-bonded heteroatoms (for example, exocyclic carbonyl groups) are not counted as heterocyclic atoms. For example, in the indole ring, the pyrrol ring is retained instead of the benzene (Scheme 13)

9. If the Number of Heteroatoms Is Equal, the Priority of Heteroatoms to Retain is N > O > S. This rule is motivated by the important role that N heterocycles play in medicinal chemistry. Sulfur has the lowest priority, because it is not able to undergo H-bonding. Therefore, in the example shown in Scheme 14, ring A of the bicyclic core scaffold of ticlopidine is retained instead of the thiophene ring B.

10. Smaller Rings are Removed First. Smaller rings are removed before larger rings.

11. For Mixed Aromatic/Nonaromatic Ring Systems, Retain Nonaromatic Rings with Priority. Aromatic systems are extremely frequent, and benzene is the most frequent ring in practically all data sets. In order to avoid too many compounds to be linked to benzene as the parent scaffold, this rule is introduced. For example, in the sertraline subscaffold shown in Scheme 15, ring B is retained and ring A is removed.

12. Remove Rings First Where the Linker Is Attached to a Ring Heteroatom at Either End of the Linker. Ring

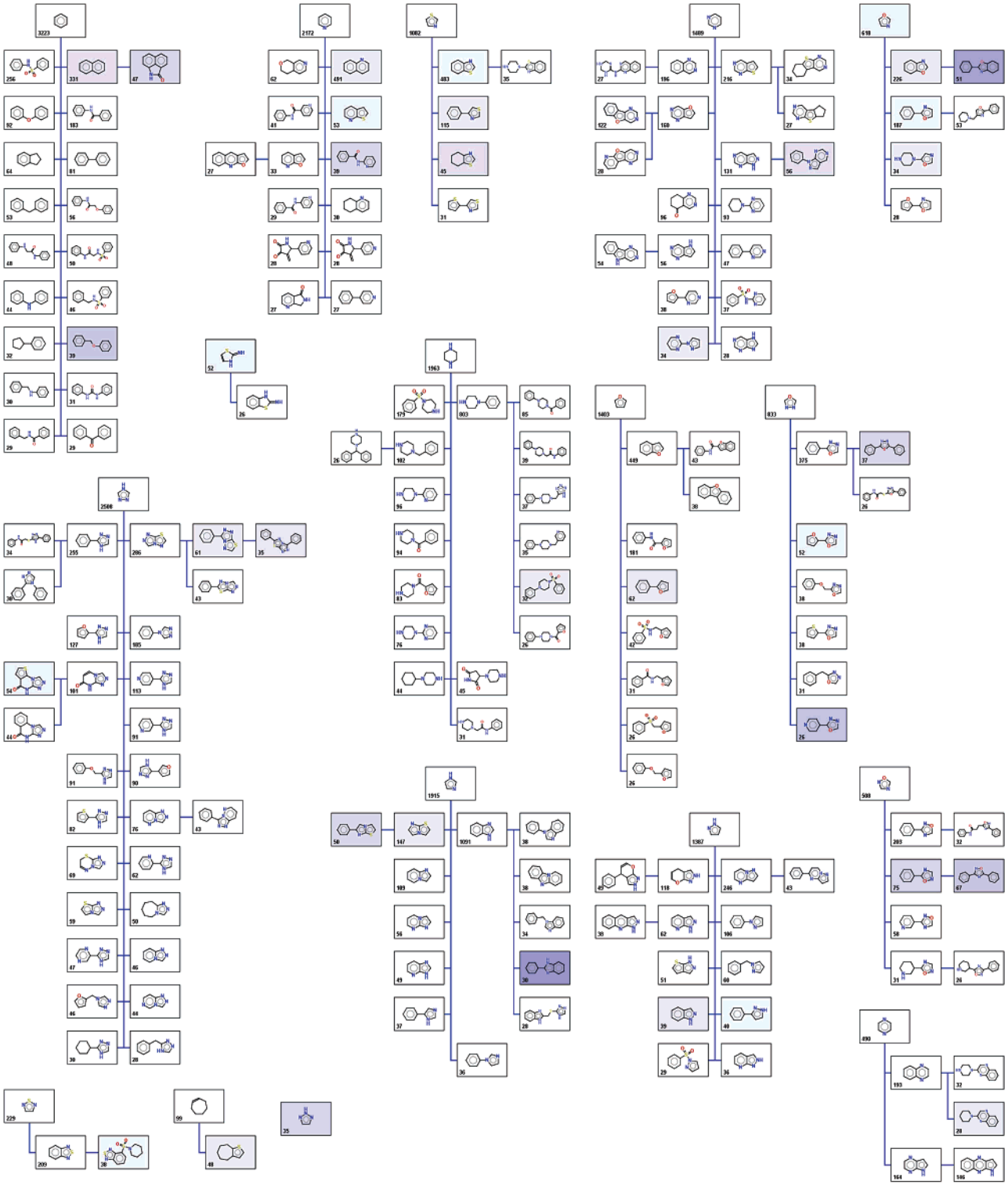


Figure 1. Scaffold tree for the results of pyruvate kinase assay. Color intensity represents the ratio of active and inactive molecules with these scaffolds. The 2-phenyl-benzooaxazole scaffold, for which the individual molecules are shown in Figure 2, can be found at the top, right corner.

heteroatoms are more easy to functionalize and, therefore, are often functionalized in the later stage of a chemical library synthesis and thus less characteristic for a chemical scaffold. For example, in the scaffold of deferasirox (Scheme 16), ring D attached to nitrogen of the triazole ring A is removed with priority.

13. Tiebreaking Rule. Remaining ties are solved by choosing from several possible remaining subscaffolds that one, whose canonical SMILES, based on the Molinspiration SMILES canonizer, has the lower rank in alphabetical order. Although the nature of this tiebreaking rule is arbitrary, the use of this rule in the classification does not mean that it will lead to a completely arbitrary overall class assignment. Consider the example of ormeloxifene (Scheme 17). After the removal of the pyrrolidine ring in the first dissection step, the scaffold ABCD is obtained. In this case, the tiebreaking rule is needed to decide whether ring D or C is to be removed. However, if ring D is removed,

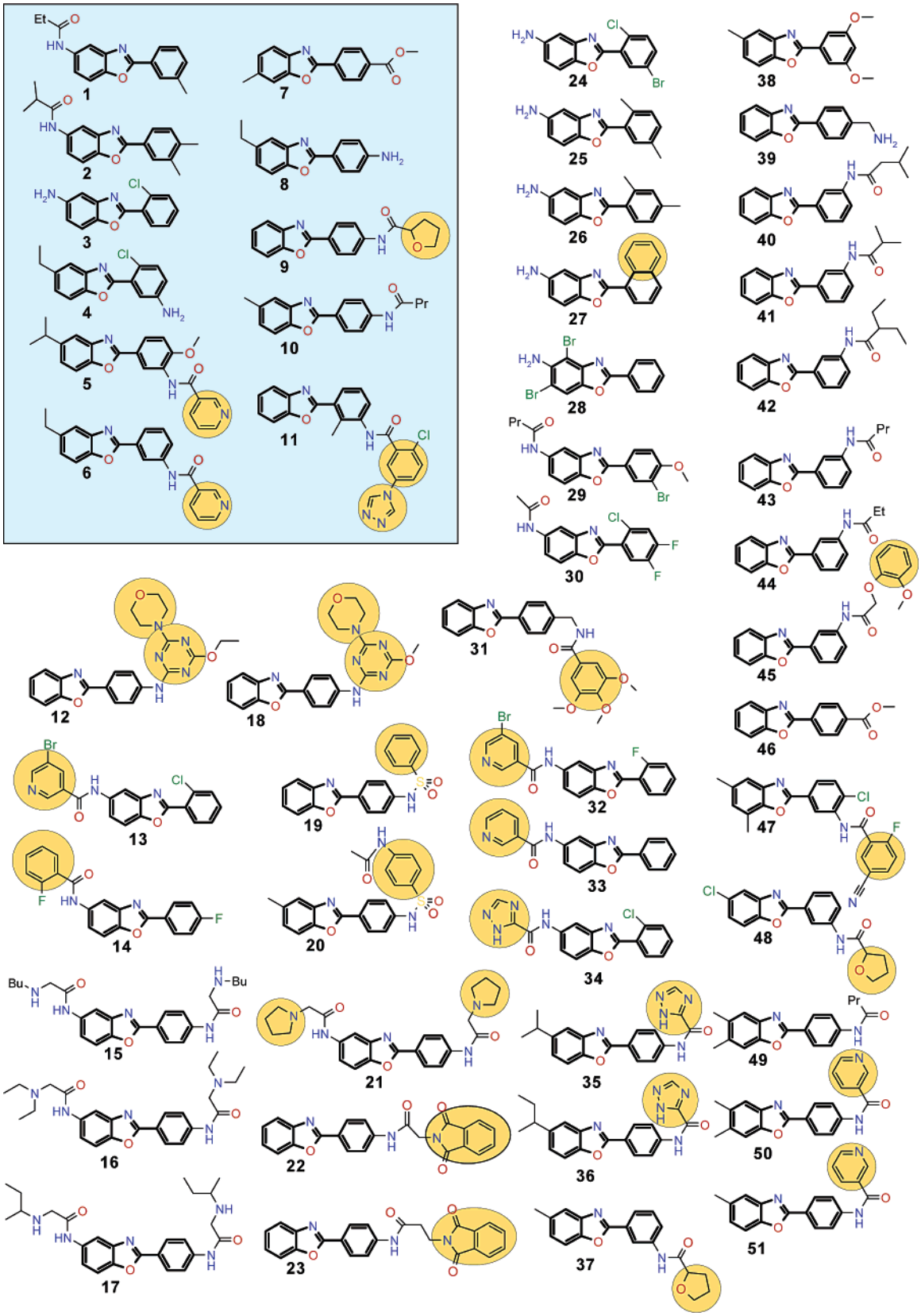


Figure 2. 2-Phenyl-benzooxazoles screened tested for activity against pyruvate kinase. Active compounds are shown in the blue box. Additional rings which would have led to a different classification when classifying by molecular framework only are highlighted in yellow.

in the subsequent step, ring C is removed, and the other way around. This means that after two dissection steps both solutions are converging again at the stage of the subscaffold AB.

CLASSIFICATION RESULTS

Before the classification results of larger data sets are presented, the whole process is once illustrated on a set of four

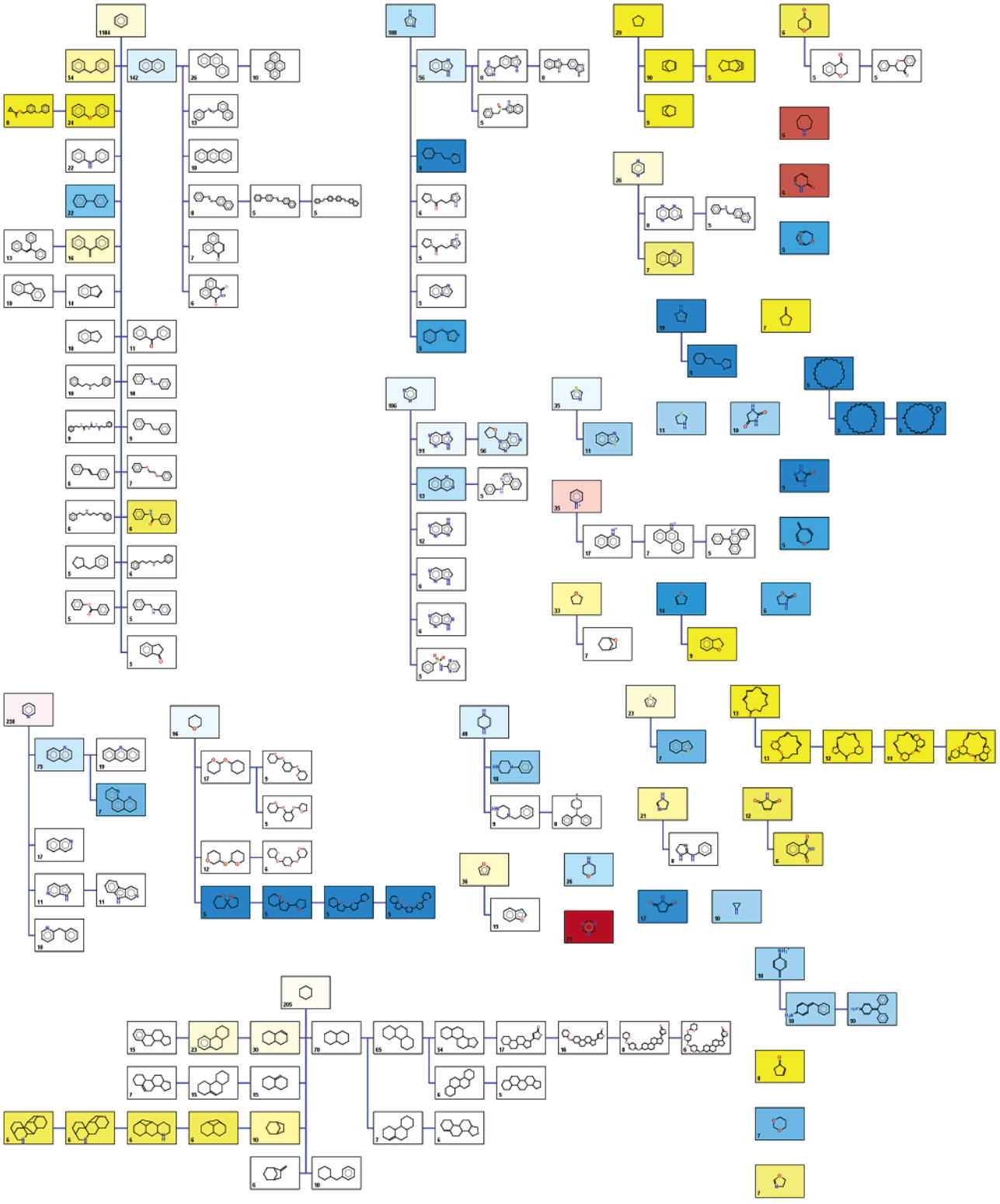


Figure 3. Scaffold tree for the pesticide data set. Fungicides are color coded by blue color, insecticides by yellow, and herbicides by red.

diazepinenones, one of the best known classes of anxiolytics: diazepam, bromazepam, zolazepam, and clotiazepam (Scheme 18). It can be seen that the molecular frameworks of these four drugs are different despite the fact that they are usually regarded as belonging to the same class of compounds. In all four cases, the linked ring is removed first according to rule 3. This already leads to the grouping of diazepam and bromazepam into the same scaffold class, whereas the other two drugs are still in their distinct classes. After the removal of the five- or six-membered aromatic ring attached to the diazepinenone ring system according to rule 6, the seven membered diazepinenone ring remaining is equal for all four molecules.

A more complex example illustrating the interplay of different rules is baccatin III (Scheme 19), which can be used as a precursor for a semisynthetic preparation of paclitaxel. According to rule 3, ring E is removed first, because this

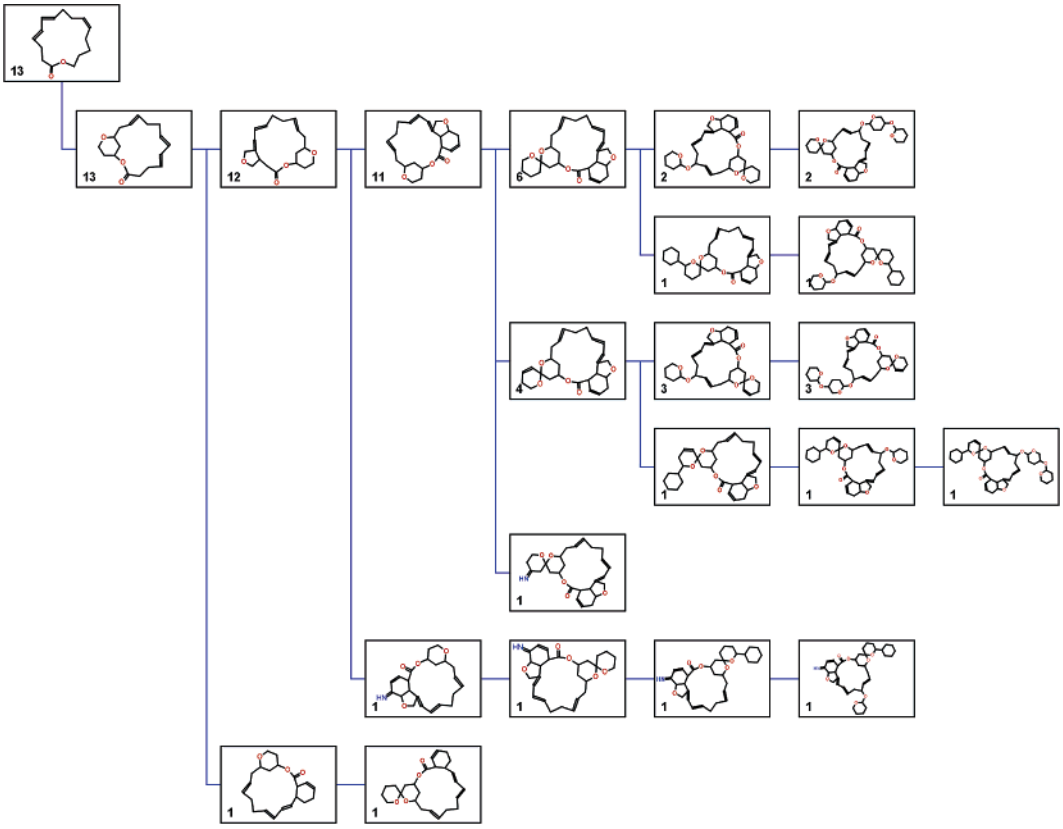


Figure 4. Scaffold tree zoom-in for the macrocyclic insecticides of the pesticide data set (Figure 3).

reduces the number of acyclic linkers to zero. This is followed by the removal of rings D and C, which retains the bridged system AB according to rule 4. Finally, according to rule 6, the eight-membered ring A is retained as a ring of unusual size. From the core, tetracyclus ABCD in both paclitaxel and baccatin III share the same scaffold hierarchy, despite their different molecular frameworks.

To demonstrate the use of the hierarchical scaffold classification in the analysis of biological screening data, we created the scaffold hierarchy trees for two data sets from the PubChem database (http://pubchem.ncbi.nlm.nih.gov/). Generated scaffold trees were displayed by using the in-house tree layout engine written in Java and SMILES molecule depictions generated by the Molinspiration toolkit.

The first example is based on the results of the pyruvate kinase assay (http://pubchem.ncbi.nlm.nih.gov/assay/assay-.cgi?aid=361), with 602 active and 50 000 inactive molecules.[18] In the hierarchy tree in Figure 1, only such scaffolds are shown which are present in at least 0.02% of the molecules in the complete data set and where at least 5% of the molecules assigned to this scaffold are active on pyruvate kinase. Color intensity is used to show the fraction of active compounds containing this scaffold. This way to visualize scaffold hierarchy is very intuitive, because color intensity coding immediately identifies those branches of the scaffold tree which contain bioactive molecules. In this case, it can be clearly seen that the activity is concentrated in a rather small number of scaffold classes, while there is a small number of active compounds in a large number of additional scaffold classes. Typically, the scaffolds with a high fraction of actives would be those analyzed first for SAR and also checked regarding their intellectual property status.

One of the scaffolds with the highest enrichment of active compounds is 2-phenyl-benzooxazole, which is found in the top, right corner of Figure 1. The individual compounds having this scaffold are shown in Figure 2. The figure shows 51 compounds in total, of which 11 are active and shown on the blue background. There are active and inactive compounds having additional rings to the 2-phenyl-benzooxazole scaffold; these rings are marked yellow. If the compounds would have been clustered only by molecular frameworks, they would have ended up in a different class. Yet, there are often pairs of compounds which are rather similar, despite one having additional rings. So compound **21** is rather similar to compounds **15**−**17** despite the additional rings. In the same way, compounds **9**, **10**, **35**, **36**, and **49**−**51** are closely related, despite different cyclic substituents. The scaffold tree helps in obtaining larger series of compounds around a common core which may be useful for deriving more robust SARs, because it is based on cyclic and acyclic side chains. In this example, some trends are visible, such as that in the meta position of the phenyl ring where an acylated amino groups is tolerated, as long as the acyl side chain is aromatic, as in compounds **5**, **6**, and **11** (with the exception of **47**), whereas nonaromatic acyl side chains such as in **37**, **40**−**44**, and **48** are not tolerated. This example illustrates that, while the scaffold tree classification as such is data-set-independent, it is beneficial to choose the hierarchy level at which to analyze the cross section through an individual scaffold tree branch depending on the data set and in such a way that the number of compounds is large enough for deriving SARs and the common core is still a reasonably large substructure of the individual molecules.

Software packages like Pipeline Pilot allow the automated generation of SAR tables listing the activity data in dependence of the side chains for a given data set based around the common core which needs to be provided by the user. Because in the scaffold tree each node is the common core for all compound structures assigned to it, the scaffold tree classification is well-suited to group compounds into subsets for the generation of SAR tables.

In a second example, three classes of pesticides have been extracted from PubChem, namely, fungicides (163 structures), herbicides (78), and insecticides (156). Additionally, a set of 5891 bioactive molecules (structures with an entry in the "Pharmacologic Action" field) was used. The resulting scaffold tree with branches containing at least 5% pesticides is shown in Figure 3. Classes of pesticides are coded by different colors, fungicides by blue, insecticides by yellow, and herbicides by red. Again, the display is very intuitive, providing, at a glance, information on which scaffolds are typical for a particular type of activity.

In both examples, the tree resolution was kept intentionally low to allow the display of a tree on a single page. But, of course, a much more detailed view is possible simply by decreasing the percentage limit for the occurrence of a branch to display. Additionally, the tree view allows also a "zoom" to more interesting areas of scaffold space. This may be exemplified for avermectins and milbemycins, products of actinomycetes from the genus *Streptomyces*, which are very potent insecticides (Figure 4, which is a more detailed view of a part of Figure 3). All the structures share a 16-membered macrocyclic lactone with fused hexahydrobenzofuran and spiroketal units. Avermectins have an attached bisoleandrosyloxy substituent at C-13, whereas that position is unsubstituted in milbemycins. These natural products are produced as a mixture of variously substituted components. The scaffold tree view allows easy navigation within these complex structural relationships.

CONCLUSIONS

With the procedure introduced here for the unique, hierarchical classification of scaffolds, we have introduced a fast, deterministic, and data-set-independent method of chemical classification. In contrast to the scaffolds used as the basis for the MEQNum,[5] each of the scaffold classes is a real chemical structure, instead of topological frameworks or reduced graphs. The method visualizes relations between different molecular frameworks, which may for example result from different members of a combinatorial chemical library having the same cyclic core and different cyclic and acyclic side chains, by tracing them to the same class at higher levels of the classification hierarchy. This makes it likely to detect chemical series in the hierarchical classification of screening hit lists. Because the method is data-set-independent, it is possible to classify compound sets individually and then overlay the sets to detect in which chemical classes there is overlap in the data sets.

The prioritization rules introduced here make it most likely that the scaffolds retained at higher hierarchy levels are chemically characteristic for the parent molecule. There is no claim made that generally the parts of the scaffolds are retained which are responsible for the biological activity of the compound class. This is not likely to be possible, because often the pharmacophoric features responsible for biologic activity can be distributed over the whole molecule including the terminal side chains. However, if a specific ring system is used in a structure class because it presents the pharmacophoric side chains in the right geometric arrangement, then this ring system may be preserved as a common core in the whole class of active compounds, although it is not the pharmacophore itself.

The method can also be used to "reverse engineer" enumerated compound collections, such as those offered by various companies selling compound libraries for screening, and identify the combinatorial library scaffolds which have most likely been used. It should also be noted, with the exception of the tiebreaking rule, that all other rules can also be easily evaluated without the use of a computer, and therefore a chemist can easily identify in which branch of the scaffold tree it would be located, especially because the set of rules is reasonably small. The possibility to color branches of the scaffold tree on the basis of the concentration of bioactivity or the presence of molecules with a specific type of activity makes this method particularly useful for presenting the analysis results to chemists. Additionally, the method allows the processing of very large data sets (we processed data sets with more than 1 million molecules), which makes this approach very useful for the analysis of results of HTS campaigns.

REFERENCES AND NOTES

(1) Markush, E. A. Pyrazolone Dye and Process of Making the Same. U.S. Patent No. 1,506,316, 1924.

(2) Böhm, H. J.; Flohr, A.; Stahl, M. Scaffold Hopping. *Drug Discovery Today: Technol.* **2004**, *1*, 217−224.

(3) Southall, N. T; Ajay. Kinase Patent Space Visualization Using Chemical Replacements. *J. Med. Chem.* **2006**, *49*, 2103−2109.

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

(5) Xu, Y. J.; Johnson, M. Using Molecular Equivalence Numbers to Visually Explore Structural Features that Distinguish Chemical Libraries. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 912−926.

(6) Cases, M.; Garcia-Serna, R.; Hettne, K.; Weeber, M.; van der Lei, J.; Boyer, S.; Mestres, J. Chemical and Biological Profiling of an Annotated Compound Library Directed to the Nuclear Receptor Family. *Curr. Top. Med. Chem.* **2005**, *5*, 763−772.

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

(8) Katritzky, A. R.; Kiely, J. S.; Hebert, N.; Chassaing, C. Definition of Templates within Combinatorial Libraries. *J. Comb. Chem.* **2000**, *2*, 2−5.

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

(10) Roberts, G.; Myatt, G. J.; Johnson, W. P.; Cross, K. P.; Blower, P. E., Jr. LeadScope: Software for Exploring Large Sets of Screening Data. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1302−1314.

(11) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.

(12) 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.

(13) Koch, M. A.; Wittenberg, L. O.; Basu, S.; Jeyaraj, D. A.; Gourzoulidou, E.; Reinecke, K.; Odermatt, A.; Waldmann, H. Compound Library Development Guided by Protein Structure Similarity Clustering and Natural Product Structure. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 16721−16726.

(14) Nören-Müller, A; Reis-Corrêa, I., Jr.; Prinz, H.; Rosenbaum, C.; Saxena, K.; Schwalbe, H. J.; Vestweber, D.; Cagna, G.; Schunk, S.; Schwarz, O.; Schiewe, H.; Waldmann, H. Discovery of Protein Phosphatase Inhibitor Classes by Biology-Oriented Synthesis. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 10606−10611.

(15) Matter, H.; Potter, T. Comparing 3D Pharmacophore Triplets and 2D Fingerprints for Selecting Diverse Compound Subsets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1211−1225.

(16) Downs, M. G.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. Review of Ring Perception Algorithms for Chemical Graphs. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 172−187.

(17) Wessjohann, L. A.; Ruijter, E.; Garcia-Rivera, D.; Brandt, W. What Can a Chemist Learn from Nature's Macrocycles? − A Brief, Conceptual View. *Mol. Diversity* **2005**, *9*, 171−186.

(18) Inglese, J.; Auld, D. S.; Jadhav, A.; Johnson, R. L.; Simeonov, A; Yasgar, A.; Zheng, W.; Austin, C. P. Quantitative High-Throughput Screening: A Titration-Based Approach that Efficiently Identifies Biological Activities in Large Chemical Libraries. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 11473−11478.

CI600338X