Quest for the Rings. In Silico Exploration of Ring Universe To Identify Novel Bioactive Heteroaromatic Scaffolds

Peter Ertl,* Stephen Jelfs, Jörg Mühlbacher, Ansgar Schuffenhauer, and Paul Selzer

Novartis Institutes for BioMedical Research, CH-4002 Basel, Switzerland

Bioactive molecules only contain a relatively limited number of unique ring types. To identify those ring properties and structural characteristics that are necessary for biological activity, a large virtual library of nearly 600 000 heteroaromatic scaffolds was created and characterized by calculated properties, including structural features, bioavailability descriptors, and quantum chemical parameters. A self-organizing neural network was used to cluster these scaffolds and to identify properties that best characterize bioactive ring systems. The analysis shows that bioactivity is very sparsely distributed within the scaffold property and structural space, forming only several relatively small, well-defined "bioactivity islands". Various possible applications of a large database of rings with calculated properties and bioactivity scores in the drug design and discovery process are discussed, including virtual screening, support for the design of combinatorial libraries, bioisosteric design, and scaffold hopping.

Introduction

Ring systems in molecules form a cornerstone of organic chemistry and therefore also medicinal chemistry and the drug discovery effort as a whole. Rings give molecules their basic shape, determine whether molecules are rigid or flexible, and keep substituents in their proper positions. In many bioactive molecules, rings are directly involved in interactions with receptors, either through heteroatoms forming hydrogen bonds with appropriate protein residues or through hydrophobic interactions. Global molecular characteristics important for the bioavailability and fate of a molecule in an organism, such as hydrophobicity or polarity, are also determined mainly by the composition of rings. Electronic ring properties including energies of frontier molecular orbitals, their distribution over the ring core, and of course, atomic charges determine the reactivity of a molecule, which in turn is responsible for its metabolic stability and toxicity. And we should also not forget the profound role of rings in cheminformatics, for example, in the large effort to identify the most efficient ring perception algorithms,[1] the hotly discussed topic of the smallest set of the smallest rings, and interesting theoretical challenges such as aromaticity and tautomerism.[2]

Ring systems also play an important role in several techniques used in modern medicinal chemistry. One example may be combinatorial chemistry and parallel synthesis, where ring systems are used as the central scaffolds of combinatorial libraries. Another popular technique applied in the drug discovery process is "scaffold hopping"[3] where the goal is to "jump" in chemistry space, i.e., to discover a new structure starting from a known active compound via the modification of the central core of this molecule.

Because of the importance of ring systems in the drug discovery process, several publications exploring this topic have appeared. Bemis and Murcko[4] analyzed 5120 known drugs to identify the most common scaffolds. They found 2506 different scaffolds, but without regard to atom type, hybridization, and bond order, half of the drugs in the database are described by just the 32 most frequently occurring scaffolds. Several authors tried to classify rings according to their characteristics. Gibson et al.[5] characterized a set of 100 aromatic rings by calculated properties that were selected because of their potential involvement in the molecular recognition of drug−receptor binding interactions. Principal components generated from these properties correlated with the in vitro biological activity of ring systems under study. Lipkus[6] presented a method for organizing ring systems based on their topology. Three simple descriptors that characterize separate aspects of ring topology were used in the study. This approach was applied to a database of 40 182 different rings that were derived from a comprehensive collection of rings extracted from the CAS registry. The study concluded that the distribution of rings is not compact but contains many significant voids. Lewell et al.[7] described the development of a drug rings database with a Web interface. The database contains ring structures from both the corporate collection and commercial databases characterized by several descriptors. A database such as this may be used in lead discovery programs where bioisosteric ring analogues are sought. Bohl et al.[8] discussed the use of shape-similarity methods to identify ring systems that are structurally similar to, and aligned with, a user-defined target ring system. This system can be used to identify alternative scaffolds for the construction of combinatorial libraries. Broughton and Watson[9] suggested a method to obtain heterocyclic ring systems suitable for use in drug design and library design, with an emphasis on the selection of systems with good adsorption, distribution, metabolism, excretion, and toxicity (ADMET) properties. This has been achieved by extraction of ring systems found in drugs that have reached phase II or later stages of drug development and launch. Principal components analysis of calculated ring properties has been used to enable the visualization of the set of heterocycles in a "useful chemical space". The authors suggest that heterocycles in successful drugs are more likely to have calculated quantities associated with lower chemical reactivity. Wilkens et al.[10] described a recursive algorithm for rapidly identifying all possible scaffolds within a set of compounds. Biological data were coupled to scaffolds by the inclusion of activity histograms, which indicate how the compounds in each scaffold class

* To whom correspondence should be addressed. Phone: +41 61 6967413. Fax: +41 61 6967416. E-mail: peter.ertl@novartis.com.

Table 1. Summary of Ring Analysis Results for Common Organic Molecules, Bioactive Molecules, and Natural Products

data set	any ring[a]	aliphatic ring[a]	aromatic ring[a]	SAS[a,b]
organic molecules	94.8	41.0	80.6	78.5
bioactive molecules	96.7	65.2	75.6	73.3
natural products	91.3	73.2	37.9	36.3

[a] The number means percentage of molecules in the particular data set containing at least one ring system of specified type. [b] SAS means simple aromatic scaffold (see the text and Figure 1 for details).

performed in previous high-throughput screening campaigns. Koch et al.[11] performed a scaffold analysis of a large collection of natural products, arranging the scaffolds in a tree-like fashion. Such an analysis, combined with annotations of biological origin and pharmacological activity, highlights the regions of chemical space explored by nature and provides guidance for the development of natural product-like compound libraries.

Cheminformatics Analysis of Ring Systems in Bioactive Molecules

The goal of the present study was to compare the properties and composition of heteroaromatic ring systems that are present in bioactive molecules with the properties of a large collection of virtual ring systems in order to identify those properties that are necessary for biological activity. The database of ring systems derived in this process would also be a good basis for various applications in the drug discovery process.

To identify ring systems typical for bioactive molecules, we analyzed the ring distribution in three large sets of molecules, namely, molecules with biological activity, average organic molecules, and for comparison, a collection of natural products. Bioactive molecules were obtained simply as a sum of structures from the World Drug Index[12] and MDDR database.[13] The "nondrug" molecules (imaging, radioprotective, or dental agents, etc.) in these databases were not included. After this step the bioactive set contained about 150 000 structures. We did not make any attempt to classify molecules according to the type of activity, target, etc.; therefore our definition of "biological activity" is very general and actually covers all molecules interacting with any of the known pharmaceutical targets. For comparison, a database containing typical organic molecules was generated from a collection of catalogues of various commercial compound providers comprising more than 6 million unique structures by selecting 150 000 molecules by a representative selection procedure implemented in PipelinePilot.[28] And finally, structures from the Dictionary of Natural Products[14] were used for the analysis of natural products. All the data were cleaned and normalized by removing molecules with apparent valence errors, by removing organometallic structures, and where possible, by neutralizing charges. All these three databases were analyzed with respect to the presence or absence of rings, and all identified ring systems (defined as assemblies of fused and spiro rings and also including attached exocyclic multiple bonds) were extracted. Throughout this article we refer to these assemblies as ring systems and scaffolds interchangeably.

Results of the analyses are summarized in Table 1, where the percentages of molecules in each of the particular data sets containing at least one aliphatic ring, aromatic ring, and simple aromatic scaffold (see below) are listed. The sum of structures containing aromatic and aliphatic rings is not 100% because one molecule may contain both aliphatic and aromatic rings. The results prove that the chemistry of bioactive molecules is clearly dominated by rings (96.7% of bioactive molecules contain rings), mostly aromatic (75.6% of bioactive molecules
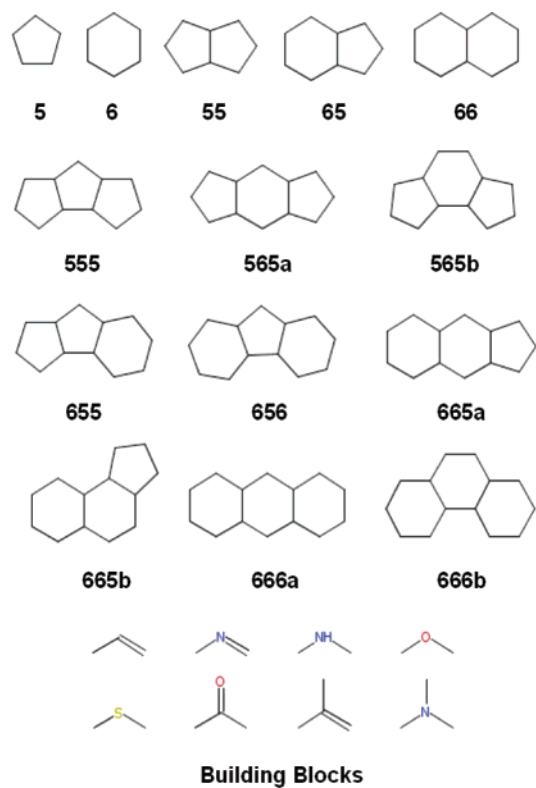
Figure 1. Types of ring systems considered in this study. The eight atom types used to construct the systems are shown at the bottom (see the text for details).

contain aromatic rings). The organic molecules have a similar ring composition, although they contain slightly more aromatic rings and fewer aliphatic rings than bioactive molecules. As expected, the situation for natural products is different; most structures in this class are based on aliphatic ring systems, and only 37.9% of natural products contain aromatic rings.

In the following analysis we will focus the study of bioactive molecules on the simple aromatic (SA) scaffolds. This is a subset of aromatic scaffolds defined as fused conjugated ring systems consisting of one to three simple five- or six-membered rings (see Figure 1). Such fused aromatic systems are very common in bioactive molecules; they are contained in 73.3% of molecules in this data set. Although modern medicinal chemistry is trying to slowly abandon the area of "flat" aromatic molecules, as may be exemplified by introduction of diversity-oriented synthesis[15] or by increasing interest in the chemistry of natural products,[16] classical drug structures are still clearly dominated by aromatic cores (see also Table 1), in some cases even inherited from the first-generation drugs originating from aromatic organic dyes.[17]

The number of unique SA scaffolds found in bioactive molecules, however, is surprisingly small. Only 780 SA scaffolds were identified within 149 437 structures. Additionally, 216 of these scaffolds are singletons (present only once in the entire bioactive collection). Only 10 such scaffolds are present in more than 1% of bioactive molecules and 64 in more than 0.1%. The most common scaffolds of this type are shown in Figure 2. Our results are consistent with the findings of Bemis and Murcko[4] about the relatively small number of unique rings in bioactive molecules. Despite this, the low diversity of simple aromatic scaffolds present in bioactive molecules is surprising. Why is this so? One can think of various possible reasons: (1) biological activity is limited to a very small portion of ring space; (2) most of the aromatic heterocyclic rings are simply

Figure 2. Most common aromatic scaffolds present in bioactive molecules.

not synthetically accessible; (3) the entire ring universe is so gigantic that the entire current synthetic effort is just scratching its surface.

Construction of a Large Virtual Scaffold Library

To learn more about the reason for the relatively low diversity of bioactive heteroaromatic scaffolds, we decided to construct all SA scaffolds and to find out whether there is a difference in properties between the subset of scaffolds found in bioactive molecules and other scaffolds. We considered the 14 relatively simple skeletons shown in Figure 1. The molecules were constructed from eight basic atom types also shown in the Figure 1, including aromatic carbon, pyridine and pyrrole nitrogens, oxygen, sulfur, and sp^2 carbon with exocyclic carbonyl bond, which is also a common component of drugs. When constructing scaffolds, we applied several simple rules to exclude "too exotic" systems, including limiting the maximum number of ring heteroatoms to 4, limiting the number of exocyclic carbonyl bonds to 2, and preventing systems with unstable bonds between O, S, and NH atoms. The correctness of all the constructed systems was checked with respect to valence rules so that any of the generated rings systems may also exist, at least theoretically, as a stand-alone molecule.

We did not perform any normalization of tautomers. One of the main goals of this study was to identify commercially available molecules containing scaffolds with high bioactivity potential, and we could not know in which tautomeric form (which of course depends also on the particular substituent pattern) the molecules (if existing) are stored in the commercial databases. We therefore considered all possible tautomers. Anyway, most of the tautomeric variability in druglike molecules is caused by −OH and −NH₂ substituents, which was not applicable in our case. Since, in the scaffolds under investigation, tautomeric variation was only possible because of ring N and NH atoms (imidazole-type tautomerism), the number of possible tautomers was not so high; only about 9% of the constructed structures were duplicated tautomeric forms of the same basic connectivity.

Another issue that needed to be addressed was aromaticity. This is a rather controversial topic in cheminformatics, with

practically each software tool handling aromaticity in a different manner and providing different answers to questions on whether a particular system is aromatic or not, using often rather complex algorithms.[2] Additionally, aromaticity is a more quantitative than qualitative property, so sometimes a single correct answer to such a question is not even possible. The classical Hückel "$4n + 2$" rule can provide some guidance, but according to this rule, 91 out of the 780 SA scaffolds found in bioactive molecules are antiaromatic, with the most prominent examples being phthalimide and phenothiazine, which form the basis for several well-known drugs. We therefore decided to keep all systems in our database independent of whether they formally fulfill the "$4n + 2$" π electron rule or not. By using quantum chemical descriptors (see below), we would expect to distinguish between aromatic and antiaromatic systems, thus making this distinction an outcome of the study rather than having to preempt a particular aromaticity model.

All ring construction and molecule manipulation were done by an in-house procedure written in Java based on a commercial cheminformatics toolkit.[18] All structures were manipulated and then stored as SMILES strings.

The number of scaffolds created in this way was quite large, totaling 580 165 (see Table 2). For comparison, the number of scaffolds of different types also identified in bioactive molecules and average organic molecules is shown. Overall, the average scaffold size was 14 atoms and the average molecular weight was 204.4.

To identify molecular characteristics that are typical for the subset of SA scaffolds found in bioactive molecules, we needed to characterize scaffolds by their calculated properties. We used several global properties including calculated logP,[22] polar surface area,[19] molecular weight, and number of atoms; simple properties characterizing composition of the scaffolds (number of heteroatoms of particular types, ratio of heteroatoms to carbon atoms, etc.); and several quantum chemical parameters characterizing electronic properties of the ring systems, including energetic parameters, charges, energies of frontier molecular orbitals, and atomic superdelocalizabilies[23] (see Table 3 for a complete list). To calculate quantum chemical parameters, 3D molecular coordinates were created from SMILES representa-

Table 2. Number of Various SA Scaffolds Generated in Silico and Found in Existing Data Sets

scaffold type[a]	virtual set	organic molecules[b]	bioactives[c]
5	66	41	36
6	134	55	34
55	1 672	53	30
65	7 793	289	229
66	8 300	145	107
555	17 244	1	1
565	54 781	42	18
655	68 633	39	20
656	63 395	107	56
665	196 404	215	142
666	161 743	157	107
all	580 165	1144	780

[a] See Figure 1 for definition of scaffold types. [b] 151 582 representative molecules selected from vendor catalogues. [c] 149 437 bioactive molecules from WDI and MDDR databases.

tions using CORINA,[20] a full geometry optimization was performed using the AM1 method in the MOPAC package,[21] and electronic parameters for the optimized geometry were extracted. The following check was also implemented to ensure that we are indeed getting parameters for the original structure. The MOPAC output file containing the optimized structure was converted to SMILES, and canonical forms of the original and final SMILES were compared. Structures for which both SMILES strings did not match (in most cases a ring opening during the geometry optimization procedure occurred) were subsequently discarded. Such a molecule breakdown happened in about 0.7% of the compounds in the entire data set. In practically all cases, the broken bond connected two nitrogen atoms, one of them on the fusion of two rings. This procedure helped to discard any potentially unstable molecules from the database. On completion, 575 776 scaffolds created in silico (780 active scaffolds have been removed from this set) were obtained, all of which were characterized by around 40 calculated parameters.

Analysis of Scaffolds by Self-Organizing Neural Networks

One of our goals was to identify parameters that separate the 780 scaffolds present in the bioactive molecules from the other scaffolds. On the basis of an initial visual analysis of the data, we could see that for some properties (for example, global molecular characteristics such as logP and some electronic parameters) there is practically no separation between active and inactive systems. For other parameters, however, one could observe a distinct clustering of active systems into particular regions of property space. Such a situation, when an observed property depends on multiple parameters in a complex, nonlinear way, is a typical case for the application of machine-learning techniques. We decided to use self-organizing neural networks, since we have good experience in applying this type of classification for processing large data sets[24] and because this method also provides an intuitive feedback on the correspondence between the input properties and the output.

The methodology of self-organizing neural networks (SONNs), often also referred to as Kohonen networks, is well-known, and a good textbook on this topic is available.[25] We will therefore provide only a basic description here. The main application area of SONNs is the simplification of complex multidimensional data. During the training process the original complex data matrix (in our cases, scaffolds characterized by normalized calculated properties) is processed and scaffolds are placed on a two-dimensional map in such a way that scaffolds with similar properties are located close to each other and dissimilar scaffolds

Table 3. Molecular Properties Calculated for the Scaffolds

|	
MW	molecular weight
natoms	number of non-hydrogen atoms
logP	calculated octanol−water partition coefficient[22]
PSA	polar surface area[19]
nX[a]	number of heteroatoms of this type
nON	number of O and N atoms
nONS	number of O, N, and S atoms
rX[a]	ratio of the number of heteroatoms to C atoms
rON	ratio of the number of O and N atoms to C atoms
rONS	ratio of the number of O, N, and S atoms to C atoms
qXmax	maximal charge on heavy atoms
qXmin	minimal charge on heavy atoms
qXavrg	average charge on heavy atoms
qHmax	maximal charge on hydrogens
qHmin	minimal charge on hydrogens
qHavrg	average charge on hydrogens
seXmax[a]	maximal electrophilic superdelocalizability on heteroatoms[23]
seXmin[a]	minimal electrophilic superdelocalizability on heteroatoms[23]
HF	heat of formation
EE	electronic energy
CCR	core−core repulsion energy
energy	total energy
DM	dipole moment
HOMO	energy of the highest occupied molecular orbital
LUMO	energy of the lowest unoccupied molecular orbital

[a] Three descriptors for O, N, and S atoms, respectively.

far away from each other. During training, no information about the bioactivity of the scaffolds is used; this information is only used at the end to calculate the "activity score" (ratio of active to inactive scaffolds) for each cell and to color the final map according to this score. This coloring allows visual assessment of the quality of the classification.

For the calculation, we used an in-house implementation of a self-organizing neural network protocol[26] written in Java. Each neuron consisted of multiple weights corresponding to particular descriptors. We applied a network with 100×100 neurons and a toroidal architecture. This size was a good compromise between the resolution of the resulting map and the computation time. A larger network size would lead to a higher resolution map (reducing the average number of compounds per neuron) but would also considerably increase the computation time because the data set contains almost 600 000 compounds (780 active and 575 776 "inactive"). In this setup each neuron initially contained about 57 scaffolds. The training process, however, redistributed scaffolds irregularly over the net, so at the end some neurons contained several hundred very similar scaffolds, while other neurons on the boundaries between dissimilar regions of property and structural space remained empty.

The final network after 200 training iterations is shown on Figure 3, where the ratio of "active" to "inactive" SA scaffolds in the cells is color-coded. This analysis shows that biological activity is very sparsely distributed in the chemical space, forming only a limited number of small, well-defined "activity islands". Indeed, on the network map we can see six small areas where a high concentration of active scaffolds is found. Representative scaffolds from each of these islands are also shown on the image.

As already mentioned, bioactivity depends on scaffold properties in a complex, nonlinear way, and the interpretation of this relationship is not straightforward. The general definition of bioactivity we used is simply too "fuzzy" to be described by a simple dependence on a property or combination of properties. Visual analysis of layers of the trained network, however, can provide some insight into the role of particular properties in separating active and inactive scaffolds. As an
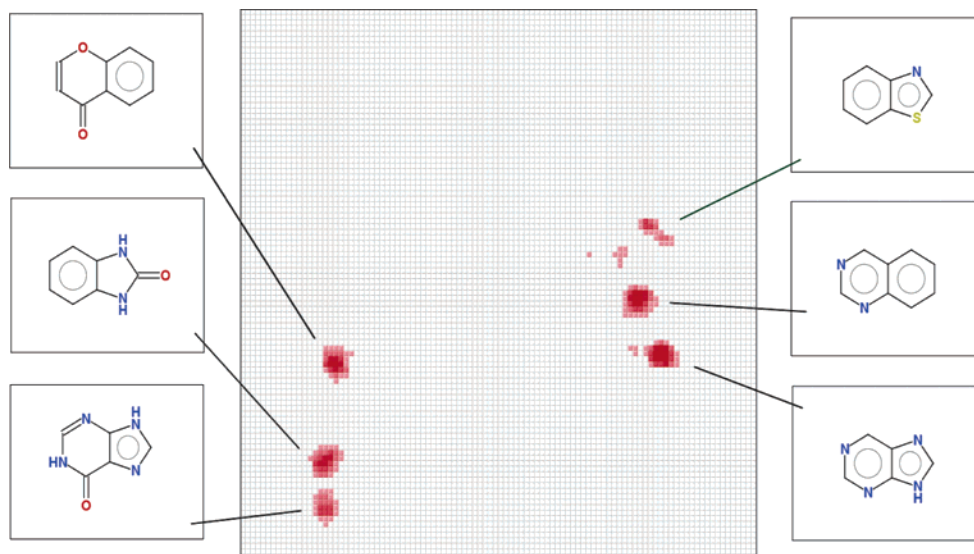
Figure 3. Trained network showing several "activity islands" and representative scaffolds for these islands.
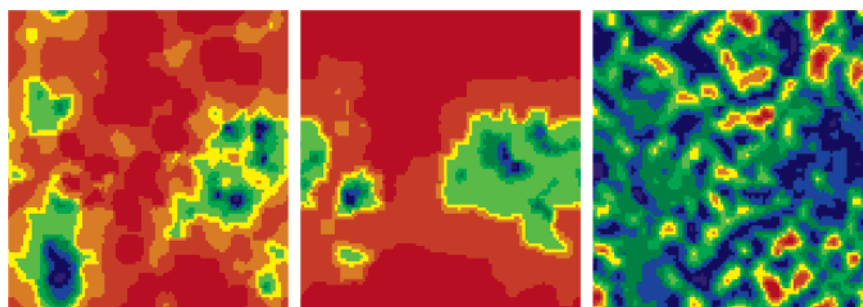
Figure 4. Network layers corresponding (from left to right) to the number of atoms, number of NOS atoms, and dipole moment. The first two layers show good similarity with the distribution of bioactivity; the distribution of dipole moment weight is independent of bioactivity.

example, three such layers are shown in Figure 4. When comparing the distribution of weights in these layers with the location of the active islands in Figure 3, one can see that the distribution of weights for the number of atoms and the number of N, O, and S atoms in scaffolds correspond well to the distribution of bioactivity while the distribution of weights for the dipole moment does not correspond to the bioactivity distribution. To get additional insight into the importance of various properties on bioactivity, we applied the Kolmogorov–Smirnov statistical test to see the difference between the distribution of particular properties for the "active" and "inactive" cells. The properties identified as the most significant were these characterizing the size of scaffolds (number of atoms, molecular weight) and their composition (number of heteroatoms of particular types), while classical global bioavailability properties (logP, PSA) were not so important. Electronic properties (the total energy and energies of frontier molecular orbitals) characterizing the stability of scaffolds and separating "classical" Hückel aromatic systems from scaffolds that violate the $4n + 2$ π-electron rule were also important, with bioactive regions corresponding mostly to the former group. We need to keep in mind, however, that such a reduction of the problem dimensionality is only an approximation and practically all properties, also including those considered less important, contribute to the classification results. For example, although the size of the scaffolds was identified as the most important parameter, with smaller scaffolds being preferred, in the active regions scaffolds consisting of three fused rings could also be found. In such cases the larger scaffold size was compensated by other beneficial properties.

Application of Results of the Scaffold Analysis in the Drug Discovery Process

In the previous section we described the successful training of a self-organizing neural network to distinguish between active and inactive ring systems. As a result, a bioactivity score may be obtained for each cell in the network as a ratio of active to inactive scaffolds in this area of chemistry space. This score characterizes the probability that the scaffolds (or molecules) located in this cell will be biologically active. In the drug design and discovery process, application possibilities of a large virtual collection of scaffolds characterized by such a score are numerous. We will discuss some examples here that have already been successfully applied at Novartis.

An apparent application area of a large database of virtual scaffolds with calculated bioactivity scores is in the purchasing of molecules for screening. Pharmaceutical companies are regularly buying large numbers of compounds from various commercial providers to extend their in-house collections and to feed high-throughput screening robots. These compounds are selected on the basis of various criteria such as novelty relative to the company collection, calculated properties, and various bioactivity models. With advantage one can also include in this process the targeted purchasing of molecules containing scaffolds with high calculated bioactivity scores that are not already present (or which are underrepresented) in the company archive. The virtual screening of the scaffolds for good ADME properties does not make too much sense here because in the actual screening samples the scaffolds will be decorated with substituents that will mostly determine the ADME characteristics of the molecules. However, the toxicity of these molecules is

another issue, and all scaffolds in our data set were virtually screened using the Novartis in-house "In Silico Tox Check" system,[26] which focuses mainly on the identification of substructures with putative carcinogenicity or mutagenecity. About 16% of rings constructed have been flagged as being potentially toxic, and these scaffolds were therefore discarded from further considerations.

Combinatorial chemistry and parallel synthesis are used routinely in the drug discovery process to automate the synthesize of large compound libraries. Central scaffolds are selected on the basis of their diversity, druglike properties, or match to a specific target. The calculated scaffold activity scores can be used to build combichem libraries around novel scaffolds with promising bioactivity potential.

Another possible application area is bioisosteric design, or scaffold hopping.[3] Generally this is a process where a target scaffold is replaced by another, sometimes considerably structually different, scaffold having, however, similar properties. This technique is often used to get molecules with better bioavailability or selectivity and molecules that are easier to synthesize or simply to replace patented structural features. A large database of scaffolds characterized by important calculated properties including bioactivity scores also provides a very good basis for the development of tools for automatic bioisosteric design. A Web tool that allows automatic identification of bioisosteric substituents and scaffolds is available to Novartis chemists on the company intranet.[27]

Last, the most challenging application of a database of virtual bioactive scaffolds is not to select scaffolds from the pool of existing compounds available for purchase but to identify and actively try to develop new synthetic methods to prepare molecules with novel active ring systems.

Conclusions

A large database of more than half a million of heteroaromatic scaffolds containing from one to three fused rings was constructed and characterized by calculated parameters describing their physicochemical properties and structural features. Self-organizing neural networks were used to identify those areas of structural and property space that are typical for scaffolds found in bioactive molecules. Probably the most interesting results of this study is a clear disproportion between a very limited number of aromatic scaffolds found in a large collection of bioactive molecules and the size of virtual scaffold space that is huge, even in the case of relatively simple conjugated scaffolds considered in this study. This clearly proves that the size of structural chemistry space is enormous and that the entire current synthetic effort is just scratching the surface. Another interesting finding is the fact that the bioactivity scaffold space is limited to several relatively small, well-defined "islands". We have to keep in mind, however, that this analysis is based only on current knowledge, and it is quite possible that there are also other active regions containing compounds that have not been tested and, even more probably, not synthesized yet. The most important properties responsible for separation of active and inactive areas of chemical space seem to be size of the scaffolds, their heteroatom composition, and their stability. Such a large database of heteroaromatic scaffolds characterized by calculated properties and by "bioactivity score" has numerous application possibilities in the drug design and discovery process.

References

(1) Downs, G. M.; 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.

(2) Roos-Kozel, B.; Jorgensen, W. Computer-assisted mechanistic evaluation of organic reactions. 2. Perception of rings, aromaticity, and tautomers. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 101−111.

(3) Böhm, H.-J.; Flohr, A.; Stahl, M. Scaffold hopping. *Drug Discovery Today: Techol.* **2004**, *1*, 217−224.

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

(5) Gibson, S.; McGuire, R.; Rees, D. C. Principal components describing biological activities and molecular diversity of heterocyclic aromatic ring fragments. *J. Med. Chem.* **1996**, *39*, 4065−4072.

(6) Lipkus, A. H. Exploring chemical rings in a simple topological-descriptor space. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 430−438.

(7) 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.

(8) Bohl, M.; Dunbar, J.; Gifford, E. M.; Heritage, T.; Wild, D. J.; Willett, P.; Wilton, D. J. Scaffold searching: automated identification of similar ring systems for the design of combinatorial libraries. *Quant. Struct.−Act. Relat.* **2003**, *21*, 590−597.

(9) Broughton, H. B.; Watson, I. A. Selection of heterocycles for drug design. *J. Mol. Graphics Modell.* **2004**, *23*, 51−58.

(10) Wilkens, S. J.; Janes, J.; Su, A. I. HierS: hierarchical scaffold clustering using topological chemical graphs. *J. Med. Chem.* **2005**, *48*, 182−193.

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

(12) WDI, Derwent World Drug Index. http://www.derwent.com/products/lr/wdi/.

(13) MDDR, MDL Drug Data Report. http://www.prous.com/product/electron/mddr.html.

(14) *Dictionary of Natural Products*, version 14.1; Chapman & Hall/CRC: London, 2005.

(15) Schreiber, S. L. Target-oriented and diversity-oriented organic synthesis in drug discovery. *Science* **2000**, *287*, 1964−1969.

(16) Breinbauer, R.; Manger, M.; Scheck, M.; Waldmann, H. Natural product guided compound library development. *Curr. Med. Chem.* **2002**, *23*, 2129−2145.

(17) Drews, J. *Quest of Tomorrow's Medicines*; Springer: New York, 2003.

(18) Molinspiration Cheminformatics Java toolkit. http://www.molinspiration.com.

(19) Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, *43*, 3714−3717.

(20) CORINA. http://www.mol-net.com.

(21) We used a modified version of MOPAC provided by Peter Bladon (Interprobe Chemical Services, Glasgow, U.K.) that is able to calculate additional atomic properties (keyword: PROPER).

(22) CLOGP. http://www.biobyte.com.

(23) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-chemical descriptors in QSAR/QSPR studies. *Chem. Rev.* **1996**, *96*, 1027−1044.

(24) Selzer, P.; Ertl, P. Identification and classification of GPCR ligands using self-organizing neural networks. *Quant. Struct.−Act. Relat.* **2005**, *24*, 270−276.

(25) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*; Wiley-VCH: Weinheim, Germany, 1999.

(26) Ertl, P.; Selzer, P.; Mühlbacher, J. Web-based cheminformatics tools deployed via corporate intranets. *Drug Discovery Today: BIOSILICO* **2004**, *2*, 201−207.

(27) Ertl, P. Cheminformatics analysis of organic substituents: identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374−380.

(28) PipelinePilot, version 5.0. http://www.scitegic.com/.