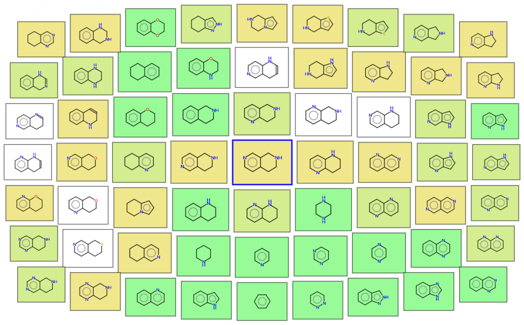
Identification of Bioisosteric Scaffolds using Scaffold Keys

Peter Ertl

Novartis Institutes for BioMedical Research, CH-4056 Basel, Switzerland
peter.ertl@novartis.com

Replacement of a central scaffold in a bioactive molecule by another scaffold with similar structural features (a procedure called sometimes "scaffold hopping") is a classical medicinal chemistry technique used to improve molecular properties and explore novel interesting areas of chemical space. The new scaffolds may be identified by database mining, match in physicochemical properties and often just by applying medicinal chemistry knowledge. In this study a novel method to find bioisosteric scaffolds is described when these are identified using similarity in simple substructure features called Scaffold Keys. Performance of the method is illustrated on several examples and a freely-available web tool https://bit.ly/scaffoldkeys allowing to find bioisosteric scaffold analogs is introduced.



Introduction

The concept of scaffold as a central part of a molecule is one of the basic concepts of medicinal chemistry. The scaffold gives a molecule its shape, determines whether the molecule is rigid or flexible and keeps substituents in their positions. Global molecular properties, such as hydrophobicity or polarity are also determined by the composition of the scaffold. Electronic properties of the scaffold (atomic charges, molecular orbitals) influence reactivity of the molecule which in turn is responsible for its metabolic stability and toxicity. The selection of molecular scaffolds and their modification, where the goal is to “jump” in chemical space and to discover a new bioactive structure with improved properties starting from a known active compound (“scaffold hopping”), is therefore an important part of the drug discovery process. Successful scaffold hop requires a lot of medicinal chemistry experience and even then, a long trial and error optimization is often needed to identify a novel scaffold with optimal balance of necessary structural features and good physicochemical properties. Computational chemistry and cheminformatics techniques can provide useful help to medicinal chemists in their effort to identify optimal scaffold replacements. Various approaches to identify bioisosteric scaffolds have been described in several good reviews [1–3] therefore it is not necessary to go into any details here.

In the present study a novel method to identify bioisosteric scaffolds is described, adding an additional tool to the medicinal chemist's toolbox. The method is based on similarity of simple scaffold structural features called Scaffold Keys and was trained to reproduce the bioisosteric scaffold replacements described in the medicinal chemistry literature.

Methodology

As mentioned in the introduction, the goal of this study was to develop a method to be able to reproduce bioisosteric scaffold pairs described in the medicinal chemistry literature. The information about the scaffold pairs was extracted from the ChEMBL database.[4] ChEMBL is an indispensable resource for medicinal chemists and cheminformaticians alike, containing in its 27th release information about 2 million molecules, 13 thousand targets and 16 million bioactivity data points extracted from 76 thousand documents. The bioisosteric scaffold pairs were identified by processing compound series described in the journal articles. Only molecules with reported activity below 10 µm in the same assay were considered and the series had to contain at least five molecules. This procedure provided 46,273 such series, most of them coming from J. Med. Chem. (18,754) followed by Bioorg. Med. Chem. Lett. (16,282) and Bioorg. Med. Chem. (4,622). The scaffolds with up to 15 non-hydrogen atoms were extracted in the same way as described in ref.[5] For all series the bioisosteric scaffold pairs (both scaffolds being connected to the identical molecule rest), were collected, providing 6470 pairs. The most frequent scaffold pairs are shown in Figure 1 including also the number of occurrences of these pairs in the literature. The pairs encoded in SMILES notation are available from the author on request. This dataset contains 3990 unique scaffolds. The most frequent ones are shown in Figure 2 as a Molecule Cloud.[6] This is a good place to reiterate our definition of the terms "simple ring", "ring system" and "scaffold" used in this study. Particularly the term scaffold is used in the medicinal chemistry literature rather freely and sometimes with ambiguous meaning. In this study these terms are used with the following meaning:

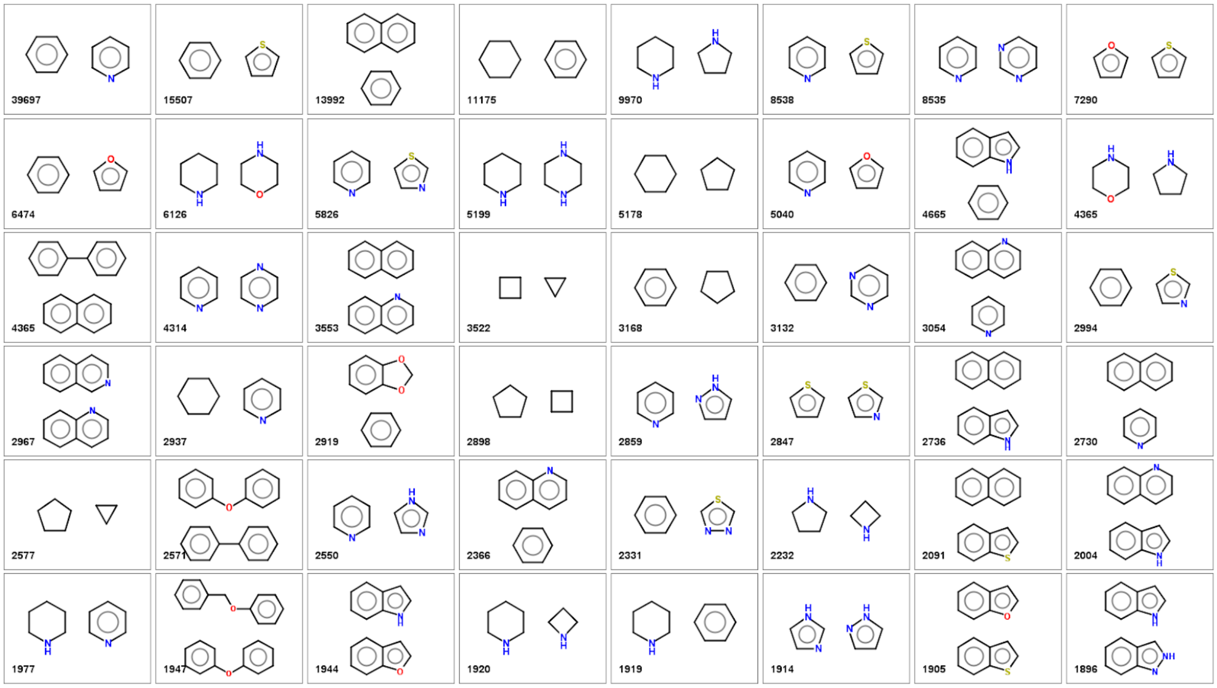


Fig. 1 The most common bioisosteric scaffolds pairs extracted from ChEMBL. The number in the corner indicates the number of occurrences of this pair in the database.

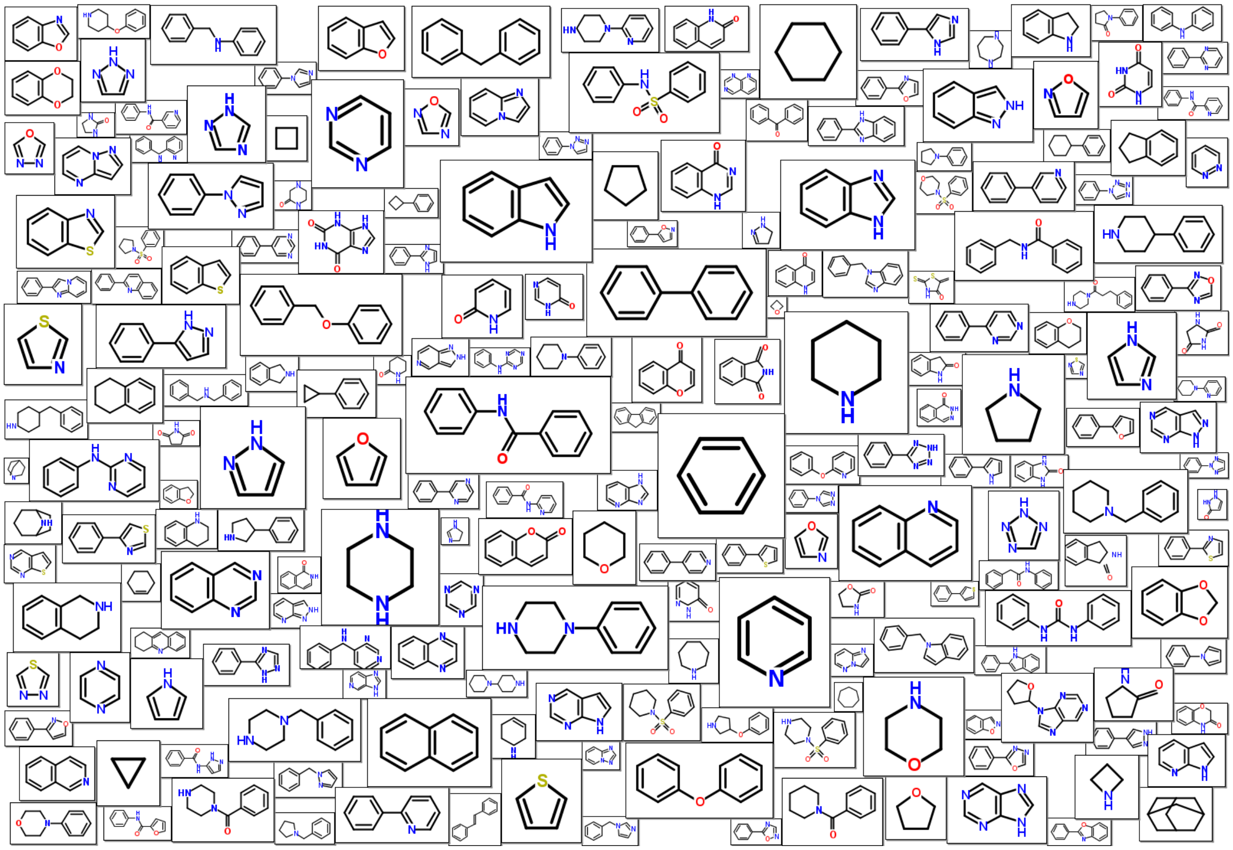


Fig. 2 The most common scaffolds present in the bioactive ChEMBL molecules displayed as a Molecule Cloud.[6]

Simple ring is a 1 ring without any exocyclic atoms.

Ring system is a single simple ring or collection of fused or spiro rings, including also exocyclic atoms connected by multiple bonds to the system.

Scaffold consists of one or more ring systems including also connections (linkers) between these systems. Exocyclic multiple bonds on rings as well as on the linkers (for example the carbonyl oxygen of an amide group connecting 2 rings) are parts of the scaffold. Non-ring substituents are not part of the scaffold. To illustrate this definition a random selection of scaffolds from our data set is shown in Figure 3, covering simple one-ring scaffolds, fused and spiro systems and also rings connected by short or longer chains.

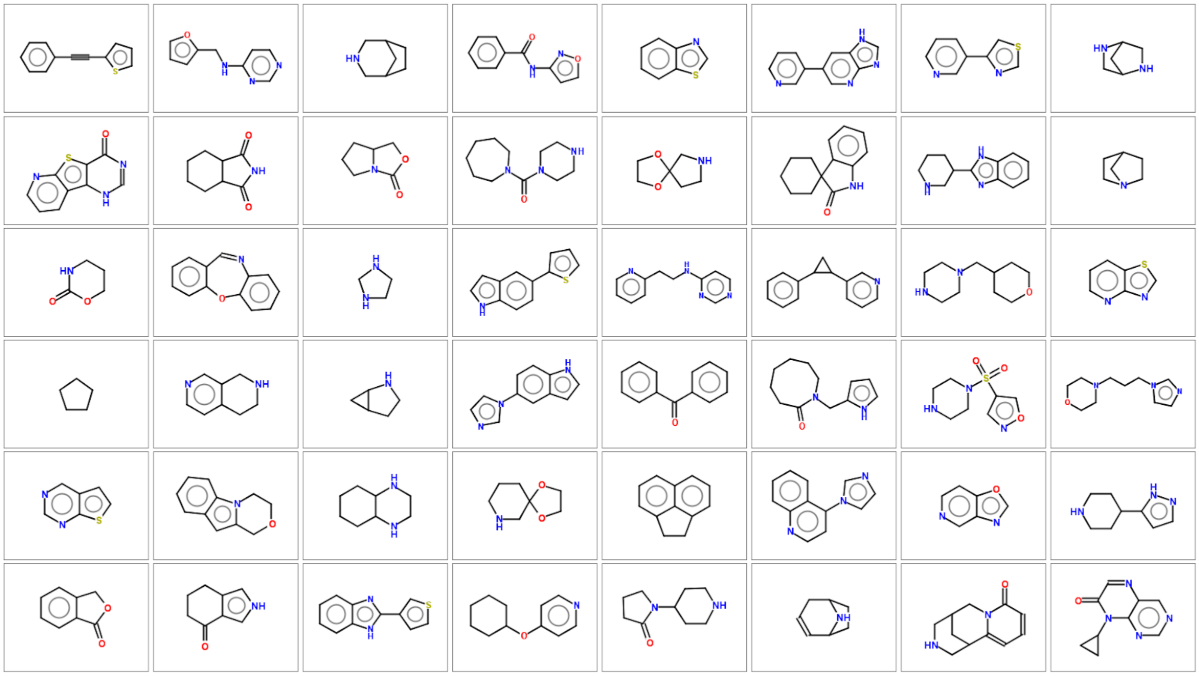


Fig. 3 Random selection of scaffolds from the training database illustrating their different types.

Results and Discussion

The bioisosteric scaffold pairs described in the previous section were collected with a goal to train a model that should be able to reproduce known and identify novel bioisosteric scaffold replacements. All experienced medicinal chemists know that the identification of bioisosteres, particularly non-classical, i.e. structurally not closely similar, is quite a challenging task. An optimal bioisosteric replacement is determined by a subtle balance of electronic, hydrophobic and steric molecular properties. Contributions of these various factors depend also on the role the replaced part is involved in. In same cases the scaffold acts as a central framework keeping the substituents in their proper 3D positions, sometimes it interacts directly with the target protein, sometimes serves only as a linker separating 2 parts of the molecule and sometimes only its physicochemical properties are important, affecting for example the solubility or hydrophobicity of the parent molecule. In the training dataset extracted from ChEMBL all these different cases are represented. One needs to be aware also of the incompleteness of the data, where many optimal bioisosteric analogs for a given scaffold are missing. The reasons for this are numerous, including unavailability of proper reagents, a fact that the proper synthetic method to access the desired analogs was not yet developed and probably the most common reason is that just nobody thought about these particular replacements.

Such a complex and incomplete data set makes the selection of a proper machine learning method and the best way to numerically characterize the scaffolds challenging. Various approaches have been applied in the past for similar tasks, including application of deep neural networks [5] or characterisation of properties of bioisosteres by quantum chemical calculations.[7] Another requirement on the potential method was, that it should be fast, to be able to suggest a large number of bioisosteric analogs for generative chemistry applications. Considering all these factors we decided at the end to use the naive Bayes classifier. The naive Bayes is a relatively simple machine learning method, but performing surprisingly well, also in situations with complex input data and the processed objects described by a limited number of simple parameters.[8] Many successful applications of naive Bayes method when applied to a broad range of cheminformatics problems have been described, including for example [9–11].

As the scaffold descriptors we used Scaffold Keys, a collection of simple substructure features described in ref.[12]. In our hands the Scaffold Keys have been shown to perform well in the diversity analysis, bioiosoteric design and mapping of chemical space.

While in the original study all keys have been used with the same equal weights, in the present work contributions of particular keys were optimized by a Bayes classifier to provide the best recovery of the literature bioisosteric scaffold pairs. The original set of keys was also enhanced by additional descriptors, including particularly more information about the scaffold topology. For the creation of the Bayes model only statistically significant data (as determined by the chi2 test) were used. The optimal set of keys was then selected by the crossvalidation experiment (1000 runs with leave 20% out) when the goal was to separate 2 sets of scaffold pairs: a set of 6470 bioisosteric pairs from literature enhanced 3990 "identity" pairs and the 200,000 random pairs created by combining the 3990 scaffolds from our set in a random manner. Once the best set of keys was obtained, the final model was derived for the whole dataset. The 33 keys providing the best performance (we term this set Scaffold Keys 2) are described in Table 1. The set includes counts of atoms and bonds of different types, various substructure features describing, branching and heteroatom environments and also descriptors describing ring and scaffold composition and their topologies.

The final model based on the Scaffold Keys 2 was able to recover 81.1 % of the bioisosteric pairs in the top 10,460 hits (the number of "true" pairs used in the training), We consider this performance very good with respect to the incomplete data set as discussed above. Among the top hits not present in the literature training data there are many scaffolds that indeed look like excellent bioisosteres, in many cases better than those in the actual training set. Some examples of the results illustrating this fact for different types of scaffolds are shown in Figure 4.

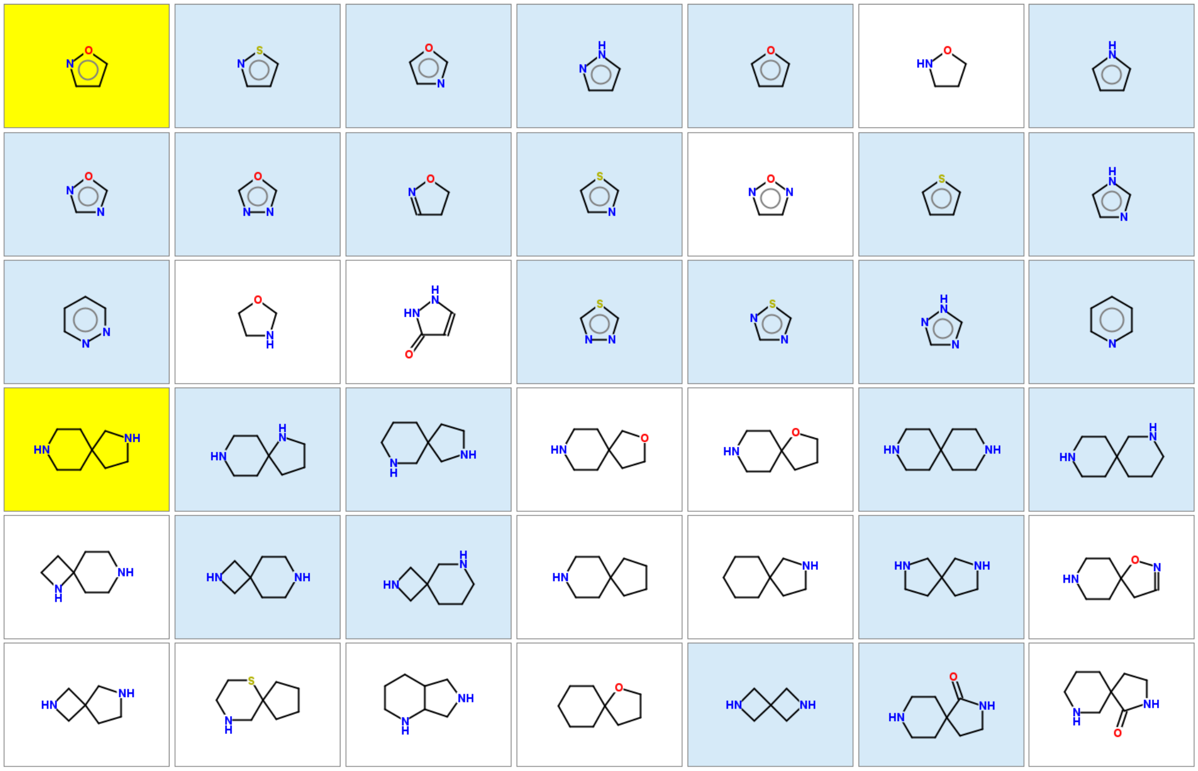


Fig 4 Example of bioisosteric analogs identified by Scaffold Keys similarity search. The query scaffolds are marked by yellow background, bioisosteric scaffolds from the literature collection have blue background.

To see whether the Bayesian approach using the Scaffold Keys 2 provides any advantage over the classical similarity search its performance was compared with a "zero model" where the analog scaffolds were identified by a standard similarity search. Exactly the same data set and performance measure were used, but the analags were identified by RDKit [13] similarity search using default parameters (Morgan fingerprints with radius 2, Tanimoto similarity). Also this procedure provides good results, the recovery of experimentally determined bioisosteric pairs is 77.7% (compared with 81.1% when using the Scaffold Keys). The major difference between the 2 methods is the fact that the Scaffold Keys procedure respects the scaffold general topology better, while the standard fingerprint-based similarity search tends to identify scaffolds with the same fragment composition, although their shape may slightly differ. Example results illustrating these differences are shown in Figures 5a and 5b.

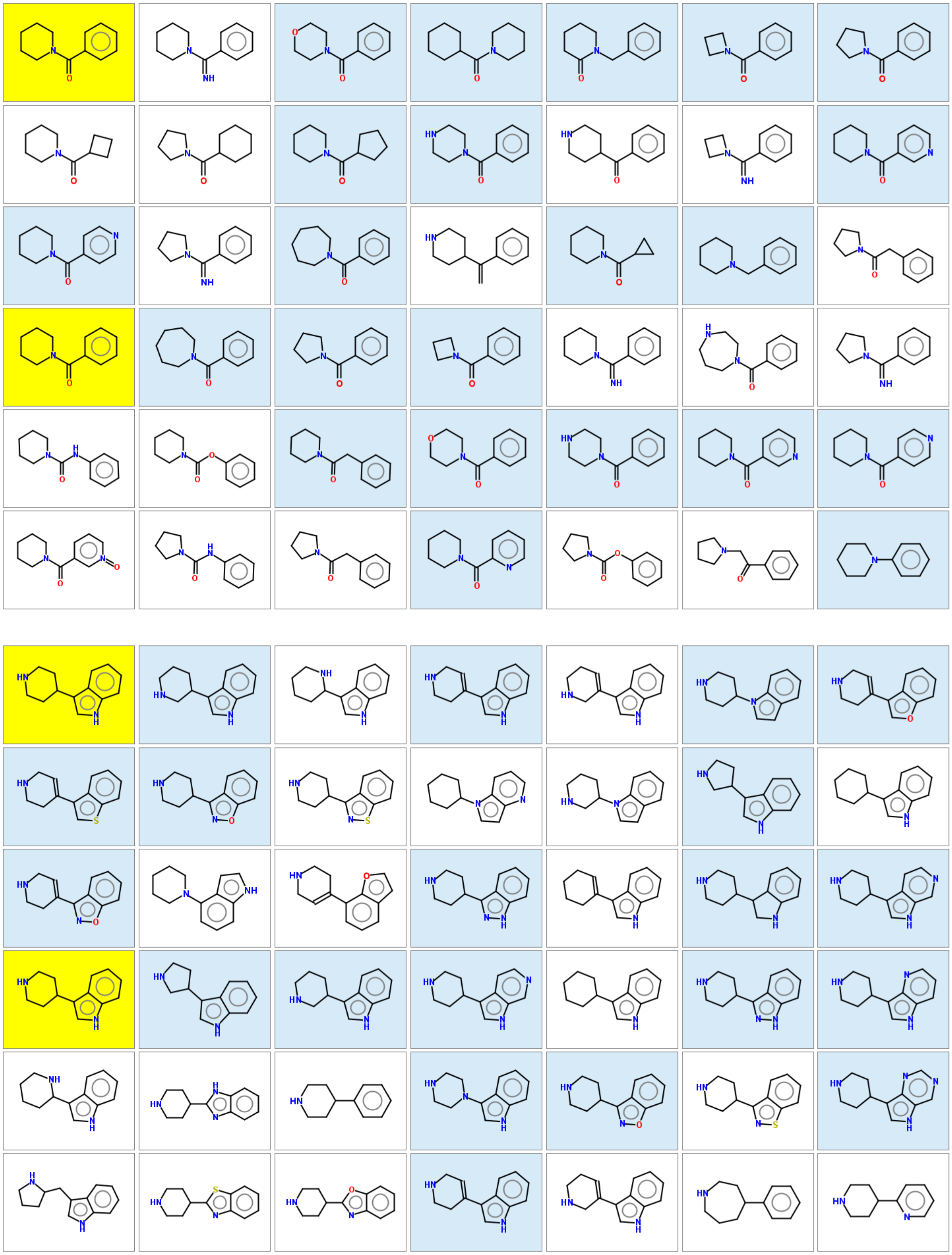
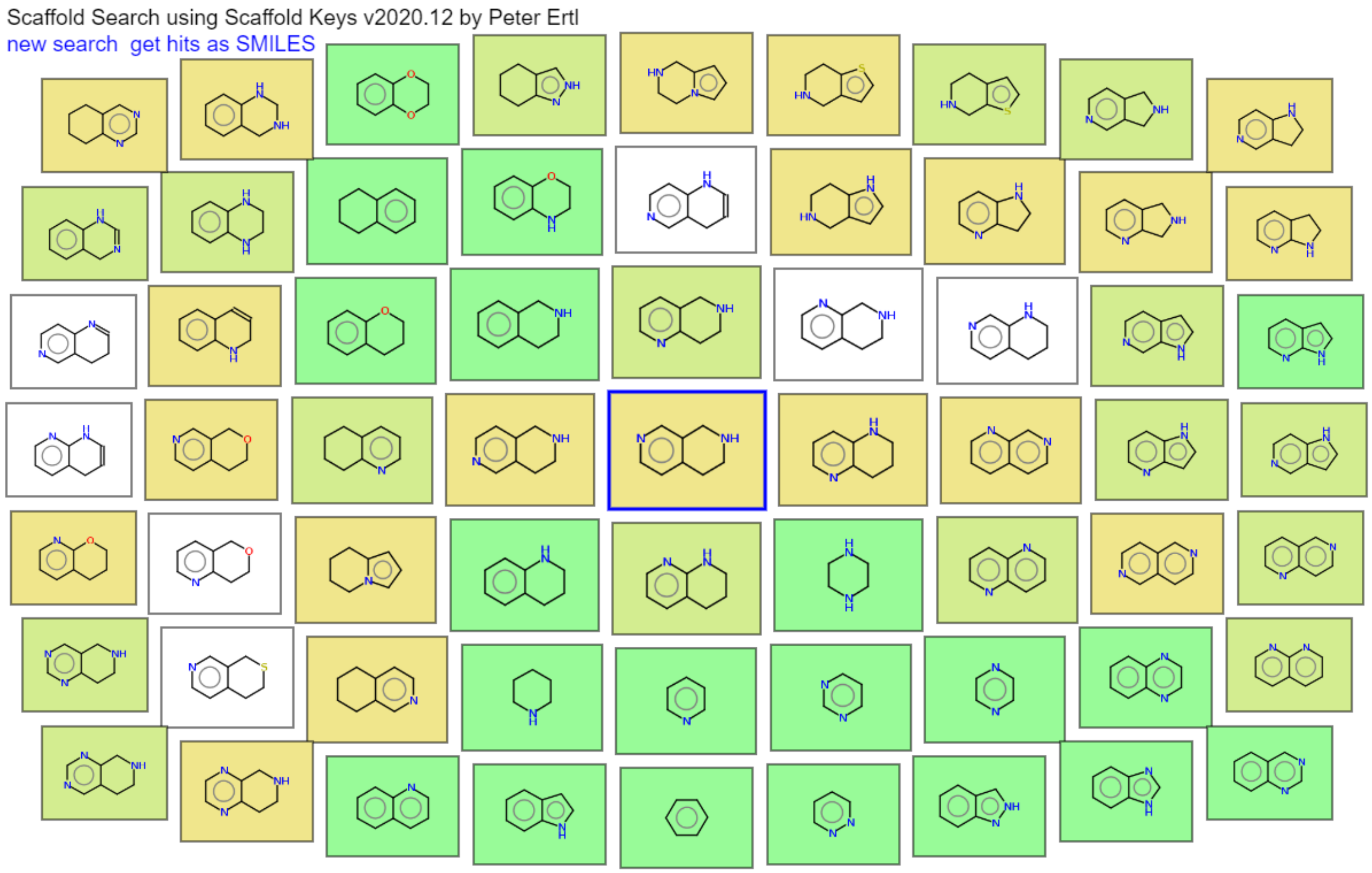


Fig. 5a,b Comparison of results obtained by similarity search using Scaffold Keys (top of the image) and the RDKit similarity search (bottom) for 2 example scaffolds. The query scaffolds are marked by yellow background, bioisosteric scaffolds from the literature collection have blue background.

Web tool for identification of bioisosteric scaffold

To offer an opportunity to identify bioisosteric scaffolds using Scaffold Keys also to the broad cheminformatics community a web tool providing this functionality was developed. A query scaffold is entered with help of the JSME JavaScript sketcher.[14] Then the query SMILES is sent to the server, where the actual search is performed in a database of more than 52,000 scaffolds extracted from the ChEMBL database characterised by their Scaffold Keys descriptors. Identified analogs are returned to the web browser where they are displayed (Figure 6). The analogs are grouped together based on their similarity and they are color coded according to their frequency in the ChEMBL database what allows chemists to focus on the more common (and therefore hopefully also better synthetically accessible) structures. The identified bioisosteric scaffolds may be also downloaded in SMILES format. Click on any scaffold in the map launches a new search with this scaffold as a query what allows an easy, interactive way to explore the huge scaffold universe and hopefully provides medicinal chemists with useful ideas for bioisosteric scaffold replacements. More detailed information about this web tool is available directly online in the tool Help page. The web tool is freely available at https://bit.ly/scaffoldkeys.



Scaffold Search using Scaffold Keys v2020.12 by Peter Ertl

Fig. 6 Output of the web tool for interactive identification of bioisosteric scaffolds. The query scaffold is in the center, scaffolds are color coded according to their frequency in the ChEMBL database.

Conclusions

A new method to identify bioisosteric scaffold analogs that is based on similarity in 33 simple substructure features (Scaffold Keys 2) weighted by a naive Bayes classifier is described. The algorithm was trained on a large set of bioisosteric pairs extracted from the medicinal chemistry literature. The method is simple, fast, may be easily implemented by any cheminformatics toolkit and provides results close to the way of thinking of experienced medicinal chemists. An easy to use web tool offering a possibility to identify bioisosteric scaffolds is available at https://bit.ly/scaffoldkeys.

Table 1 Scaffold Keys 2 - simple topological descriptors used in this study.

#	**Key description**
atom properties	
1	number of atoms in conjugated rings
2	number of atoms not in conjugated rings
3	number atoms in chains (not counting double-connected exo-chain atoms)
4	number of exocyclic atoms (connected by multiple bonds to a ring)
5	number of nitrogen atoms
6	number of nitrogen ring atoms
7	number of oxygen ring atoms
8	number of S atoms
9	number of heteroatoms
10	number of spiro atoms
11	number of heteroatoms with more than 2 connections
12	number of carbon atoms connected to at least 2 heteroatoms
13	number of atoms where at least 2 connected atoms have more than 2 connections
14	absolute value of the scaffold formal charge
bond properties	
15	number of bonds
16	number of multiple, nonconjugated rings bonds
17	number of bonds connecting 2 heteroatoms
18	number of carbon-carbon bonds when each carbon contains at least one heteroatom
19	number of bonds with at least 3 connections on both its atoms
20	number of exocyclic bonds where a ring atom is carbon
21	number of non-ring bonds connecting 2 rings, one of them conjugated and one non-conjugated
22	number of bonds where both its atoms have at least one neighbor (not counting the bond atoms) with more than 2 connections
properties of simple rings	
23	size of the largest ring
24	number of rings with more than 6 atoms
25	number of simple rings with no heteroatoms
26	number of simple rings with 1 heteroatom
27	number of simple rings with 3 heteroatoms
28	number of simple non-conjugated rings with 5 atoms
29	number of simple non-conjugated rings with 6 atoms
properties of ring systems	
30	number of ring systems
31	number of rings systems with 2 non-conjugated simple rings
32	number of rings systems with 3 conjugated simple rings
33	number of rings systems with 3 non-conjugated simple rings

Availability of data

The web tool described in this article is freely available at https://bit.ly/scaffoldkeys. The list of 6,470 bioisosteric scaffold pairs extracted from the literature and used to train the model is available from the author on request.

References

1. Langdon SR, Ertl P, Brown N (2010) Bioisosteric Replacement and Scaffold Hopping in Lead Generation and Optimization. Mol Inform 29:366–385. https://doi.org/10.1002/minf.201000019
2. Hu Y, Stumpfe D, Bajorath J (2017) Recent Advances in Scaffold Hopping: Miniperspective. J Med Chem 60:1238–1246. https://doi.org/10.1021/acs.jmedchem.6b01437
3. Bajorath J (2017) Computational scaffold hopping: cornerstone for the future of drug design? Future Med Chem 9:629–631. https://doi.org/10.4155/fmc-2017-0043
4. Mendez D, Gaulton A, Bento AP, et al (2019) ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res 47:D930–D940. https://doi.org/10.1093/nar/gky1075
5. Ertl P (2020) Identification of Bioisosteric Substituents by a Deep Neural Network. J Chem Inf Model 60:3369–3375. https://doi.org/10.1021/acs.jcim.0c00290
6. Ertl, Peter (2012) The Molecule Cloud - compact visualization of large collections of molecules | Journal of Cheminformatics | Full Text. 4:12:
7. Ertl P (1997) Simple Quantum Chemical Parameters as an Alternative to the Hammett Sigma Constants in QSAR Studies. Quant Struct-Act Relatsh 16:377–382. https://doi.org/10.1002/qsar.19970160505
8. Zhang H (2004) The Optimality of Naive Bayes
9. Bender A (2011) Bayesian Methods in Virtual Screening and Chemical Biology. In: Bajorath J (ed) Chemoinformatics and Computational Chemical Biology. Humana Press, Totowa, NJ, pp 175–196
10. Watson P (2008) Naïve Bayes Classification Using 2D Pharmacophore Feature Triplet Vectors. J Chem Inf Model 48:166–178. https://doi.org/10.1021/ci7003253
11. Townsend JA, Glen RC, Mussa HY (2012) Note on Naive Bayes Based on Binary Descriptors in Cheminformatics. J Chem Inf Model 52:2494–2500. https://doi.org/10.1021/ci200303m
12. Ertl P (2014) Intuitive ordering of scaffolds and scaffold similarity searching using scaffold keys. J Chem Inf Model 54:1617–1622. https://doi.org/10.1021/ci5001983
13. Landrum G RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. https://www.rdkit.org/
14. Bienfait B, Ertl P (2013) JSME: a free molecule editor in JavaScript. J Cheminformatics 5:24. https://doi.org/10.1186/1758-2946-5-24