DOI: 10.1002/cmdc.200800178

On the Art of Compiling and Using 'Drug-Like' Chemical Fragment Spaces

Jörg Degen,[a] Christof Wegscheid-Gerlach,[b] Andrea Zaliani,[a] and Matthias Rarey*[a]

Ever since the first rational approaches to the discovery of promising lead candidate structures were applied, it has been a challenge for both medicinal and computational chemists to assess, generate, and combine promising structural motifs to form new and potent chemical entities for biological screening against potential drug targets. Many scientists have committed themselves to the analysis and identification of valuable chemical building blocks and have also developed strategies on how to best recombine them. In this context, the retrosynthetic fragmentation and recombination of chemical motifs derived from known inhibitors is a common and well-known procedure. Meanwhile, fragment-based approaches have become established and valuable processes in pharmaceutical lead discovery and validation. Several application studies have yielded promising lead candidates.[2]

Chemical space is huge. Corporate as well as public databases are in the millions and are still increasing in size in order to cover a larger part of the chemical universe. For several good reasons, there is the common trend to standardize experimental and computational protocols in pharmaceutical research. This trend demands systematic and consistent approaches, although they can hardly match the creativity and intuition of medicinal chemists. Consequently, they can and should not substitute, but rather assist, the expert in this task. The most prominent automated example for fragment generation is the retrosynthetic combinatorial analysis procedure (RECAP).[3] It was the first of its kind to apply 11 distinct rules that were supposed to model chemical motifs that could easily be formed by combinatorial chemistry. In this context, the "fragment space" concept was introduced. In contrast to a fragment library, such a space consists not only of a set of fragments, but also of a set of rules that specifies how to recombine fragments by fusing the respective chemical motifs.

RECAP is widely used and often referred to, yet even though authors frequently state to have used modified improved versions of the original, actual publications that communicate the extensions that were carried out are rare. An extension of the fragment space concept was recently published, but with a focus on obtaining scaffolds and not on retaining supposedly 'drug-like' substituents or functional groups.[4] Apart from that, the question remains what a 'drug-like' fragment space actually is, and whether or not 'drug-likeness' depends on the origin of the fragments: that is, if they necessarily have to be derived from drugs. In this context, it is highly interesting and important to measure the extent and accuracy with which current models and methods are able to represent the available chemical space.

In an attempt to improve existing approaches for the automatic decomposition of molecules into fragments, we compiled a new and more elaborate set of rules for the breaking of retrosynthetically interesting chemical substructures (BRICS) and used this for obtaining fragments from biologically active compounds and vendor catalogue sources. Based on this, we compiled corresponding fragment spaces by specifying a complementary set of rules for the recombination of the corresponding chemical motifs. Furthermore, we put considerable effort into compiling a set of high-quality, high-performance, and, in contrast to all other approaches, publicly available fragments that are meant to serve as a possible basis for various molecular design objectives and techniques.[1] We incorporated more elaborate medicinal chemistry concepts and, for example, modeled explicit isosteric replacements for cyclic and acyclic cases and further distinguished activated from inactivated heterocyclic ring systems and their corresponding substituents. Overall, this work led us to more comprehensive sets of fragments, and the corresponding fragment spaces show a significant increase in performance over existing methods. Moreover, by incorporating fragments from vendor catalogue sources, the performance can be increased even further.

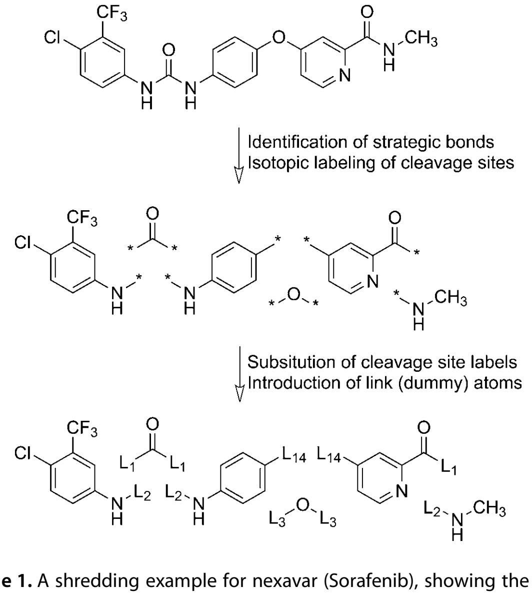
The shredding procedure we used for BRICS applies all possible retrosynthetic cuts simultaneously, which avoids the generation of overlapping (redundant) fragments. This is in accordance with RECAP and simplifies the calculation later on. Scheme 1 shows a simple fragmentation example and highlights the key steps. In addition to splitting retrosynthetically relevant bonds, we directly included substructure filters into the shredding procedure to avoid the generation of unwanted chemical motifs as well as small terminal fragments such as single hydrogen and halogen atoms, hydroxy, nitro, carboxylate, methoxy, methyl, ethyl, and isopropyl groups. These motifs are therefore discarded or left uncleaved, respectively.

The BRICS model consists of 16 chemical environments indicated by link atoms of different types. The corresponding fragment prototypes are depicted in Scheme 2 and show only the direct chemical environment of the cleavage sites for reasons of simplicity. Therefore, the diversity of the fragments is within the R groups that can also contain further links. Note that the carbonyl and alkyl fragments are shown twice (L_1/L_6, L_4/L_8). This is because we wanted to keep track of their origin for medicinal chemistry and modeling reasons, that is, whether they appeared as cyclic or acyclic substituents or linkers. The corresponding fragment space results from the definition of the

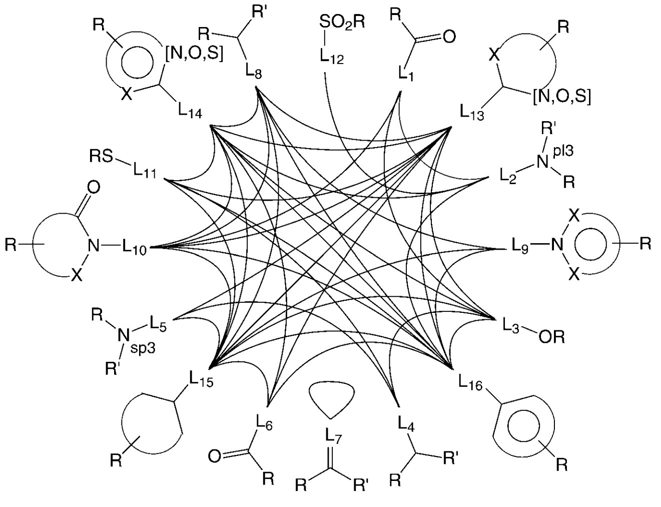
[a] J. Degen, Dr. A. Zaliani, Prof. Dr. M. Rarey
Center for Bioinformatics, University of Hamburg
Bundesstrasse 43, 20146 Hamburg (Germany)
Fax: (+49) 40-42838-7352
E-mail: rarey@zbh.uni-hamburg.de

[b] Dr. C. Wegscheid-Gerlach
Medicinal Chemistry VII Computational Chemistry
Bayer Schering Pharma AG, Müllerstrasse 178, 13342 Berlin (Germany)

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/cmdc.200800178 and from the authors' website[1].


Identification of strategic bonds
Isotopic labeling of cleavage sites
Subsitution of cleavage site labels
Introduction of link (dummy) atoms

Scheme 1. A shredding example for nexavar (Sorafenib), showing the key steps in the generation of the fragments.



Scheme 2. Fragment prototypes used in BRICS. Each chemical environment is represented by a so-called link (dummy) atom of a certain type. The diversity of the fragments is in the R groups that can contain additional link atoms (or be a link atom themselves). The R′ groups can also consist of hydrogen only. Circle patches indicate rings of various sizes (possibly including annealed or bridged ring systems as well), and the atom label 'X' stands for any of the elements C, N, O, or S. Every line between any two fragments indicates that these can be connected by forming a new bond between the atoms adjacent to the links and by deleting the links themselves. The complete fragment space including the user-customizable compatibility specification and all input data files is available at our BRICS webpage.[1]

compatibility of the respective chemical environments (indicated by lines connecting the link atoms).

We implemented the new set of rules in SMARTS notation[5] and used the shredding functionality of the Recore program for retrosynthetic fragmentation, which works on 3D molecular structures and thereby retains the input coordinates in the fragments.[6] This setup, in particular, easily enables potential users to adapt and modify the set of rules for splitting and recombination. For reference purposes, we also re-implemented the RECAP rules from the original publication to the very best of our knowledge into the same shredding framework. Both sets of rules were applied to the World Drug Index (WDI)[7] and the 'drug-like' subset of the ZINC database.[8,9] We pre-processed and pre-filtered the two catalogues using a custom PipelinePilot protocol[10] to remove ions, metals, and unwanted chemical motifs and to obtain consistent protonation states. For further use with Recore we generated low-energy conformations by using Corina.[11] Note that the RECAP results may differ with respect to the original publication, a result of the different data sources used, the additional filtering steps, and the comprehensive re-implementation carried out.

The results of the shredding are shown in Tables 1 and 2. In general, BRICS is able to cleave about 10% more molecules than RECAP. At the same time, the number of fragments with more than one connection point is higher, which leads to more branching possibilities in the end, that is, a greater number of possible topologies during the construction of molecules that can be directly observed in the results of the performance evaluation.

Table 1. Results of the shredding procedure for the World Drug Index (WDI).

WDI Database[a]	BRICS	RECAP[b]
Uncleaved compounds	20715 (35%)	28362 (47%)
Cleaved compounds	39163 (65%)	31516 (53%)
Unique fragments	18291 (100%)	15650 (100%)
1-connection fragments	13256 (72%)	12236 (78%)
2-connection fragments	4344 (24%)	3079 (19%)
3-connection fragments	591 (03%)	290 (02%)
4-connection fragments	83 (<1%)	36 (<1%)
5-connection fragments	12 (<1%)	7 (<1%)
6-connection fragments	5 (<1%)	1 (<1%)
7-connection fragments	–	1 (<1%)

[a] We used a pre-filtered collection of compounds derived from the 2004 version of the WDI database containing ~60000 compounds of pharmaceutical interest. [b] The results are derived with the reference re-implementation; see text for details.

Table 2. Results of the shredding procedure for the 'drug-like' subset of the ZINC database.

ZINC Database[a]	BRICS	RECAP[b]
Uncleaved compounds	214793 (11%)	415762 (21%)
Cleaved compounds	1806262 (89%)	1605293 (79%)
Unique fragments	93309 (100%)	101889 (100%)
1-connection fragments	76196 (81%)	90023 (88%)
2-connection fragments	15964 (17%)	11395 (11%)
3-connection fragments	1118 (01%)	453 (<1%)
4-connection fragments	30 (<1%)	17(<1%)
5-connection fragments	1 (<1%)	1 (<1%)

[a] We used a pre-filtered collection of compounds derived from the 'drug-like' subset of the 2006 version of the ZINC database containing ~2.1×10^6 compounds from various vendor catalogue sources. [b] The results are derived with the reference re-implementation; see text for details.

The first step we took for compiling a high-performance and publicly available fragment space was to determine the overlap between the shredding results for the WDI and ZINC databases. Therefore, we identified structurally identical fragments by using canonical SMILES strings.[12] We did this separately for the BRICS and for the RECAP results to assess and compare the performance of the corresponding fragment sets. The resulting fragment collections contain 4800 (BRICS_4k) and 4125 fragments (RECAP_4k), and will be called the 'intersection' in the following. This shows that the BRICS approach increases the amount of overlapping fragments between WDI and ZINC, and thus simultaneously provides a more 'drug-like' and more general fragment data source.

In a second step, we estimated the performance of both intersections. Therefore, we generated different query sets by compiling three diverse random collections of molecules. The first two sets were compiled from the collections also used for shredding, that is, the pre-filtered WDI and ZINC databases. The third collection was derived from the identically pre-processed and filtered PubChem database.[13] This is a 'true' test set because we used only entries that are contained neither in ZINC nor in the WDI. Each individual set consists of 30 000–35 000 compounds that have at least one strategic bond, which would be cleaved during the fragment generation according to the corresponding set of rules. After having compiled the query sets we then tried to reconstruct each query molecule out of the corresponding fragment sets by compiling a corresponding fragment space and applying the same rules for recombination that were used for the cleavage before.

Because these spaces are combinatorial in nature and typically contain between 10^{13} and 10^{19} molecules of reasonable size, classical sequential comparison cannot be applied. Instead, the Feature Tree fragment space search algorithm was used for this task, as it is the only method available that is able to find the globally most similar compounds in a fragment space of this size in reasonable time by employing a truly combinatorial optimization scheme.[14–16] From the calculations, we obtained the 25 most similar results for each query. Because the Feature Tree is a topological descriptor and we wanted to consider more structural details, we re-ranked every individual solution set by using the MDL "public keys"[17] and calculating pairwise Tanimoto distances.[18] For further analyses, we kept only the most similar solution to each query as a single result. This type of procedure is conceptually quite similar to the one used by Mauser and Stahl.[4]

From the results of the calculations, we generated histograms of the similarity values for the best solution that was generated for each query molecule. The normalized distributions contained in Figure 1 clearly show that we could significantly improve the RECAP performance for all test cases by using the new BRICS model, that is, the corresponding fragments and the respective set of rules for recombination. This led us to the conclusion that the fragments produced by BRICS seem to be more general in nature and lead to a greater number of reconstructed queries as well as solutions with higher overall similarity values in case the original queries could not be exactly rebuilt. This holds not only for 'drug-like' molecules contained in the WDI query set, but also for compounds from vendor catalogues contained in ZINC and molecules taken from PubChem.

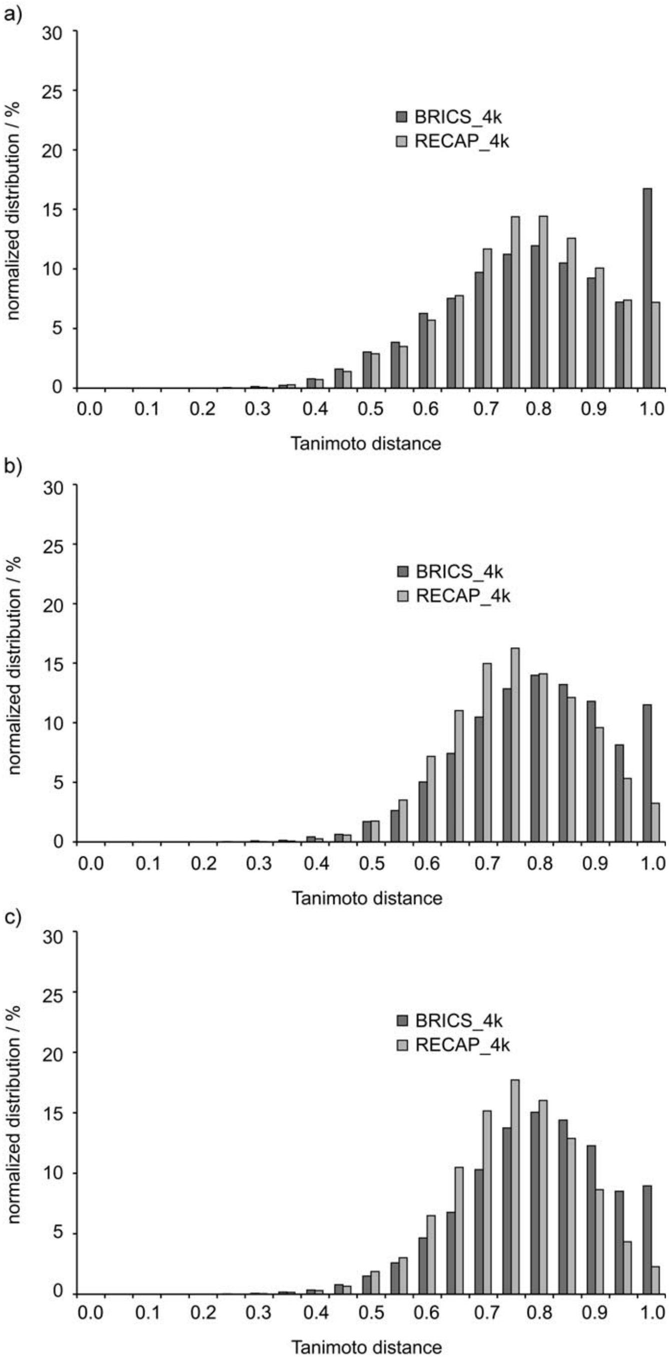


Figure 1. Comparisons of the performance of the BRICS and RECAP intersections for the various query sets (a) WDI, b) ZINC, and c) PubChem). The histograms show the normalized distribution of the similarity values (Tanimoto distance of the respective MDL public keys) of the closest solution that could be constructed out of the corresponding fragment set with respect to each query molecule. See text for details.

Because the BRICS intersection contains roughly 20 % more fragments than the RECAP intersection, the question arises whether or not the fragment count alone has an influence on the results. To account for this, we created several reduced

BRICS spaces of the same size as the RECAP space and carried out the same analyses as described above. It clearly turned out that a reduction by this order of magnitude does not have a significant influence on the results. Therefore, the extension of the RECAP approach seems to be well justified, and the corresponding shredding specification and connection rules seem to better model the chemical motifs contained in the various types of existing compounds.

To further improve the performance of the BRICS intersection, we enriched the set by adding all ZINC fragments that have a certain minimum similarity to the remaining WDI fragments. Therefore, we calculated all pairwise Tanimoto distances of the corresponding MDL "public keys" in binned ranges of molecular weight (30 Da) and chose a similarity of 0.9 and 0.8 as thresholds, which are reasonable values for still obtaining significant structural similarity.[19] By doing so, we generated two new fragment sets containing about 9300 (BRICS_9k) and 22 000 (BRICS_20k) fragments (Table 3). The RECAP intersection was not followed up owing to the huge differences in the performance that we already observed with the intersection alone.

Table 3. Summary of the fragment collections compiled.

	BRICS[a]	RECAP[b]
Intersection	4800	4125
Intersection enriched 0.9 (BRICS_9k)	9344	–
Intersection enriched 0.8 (BRICS_20k)	22343	–

[a] The enriched intersections were generated by additionally taking all ZINC fragments that have at least the given similarity value (Tanimoto distance using MDL public keys in binned molecular weight ranges) to any of the WDI fragments. [b] The RECAP intersection was not followed up; see text for details.

We then used the same computational setup as before to estimate the performance of the new sets. The results are depicted in Figure 2 and show another significant increase in the performance with respect to all test sets. For the WDI queries this improvement is smallest, but the results are already very good with the intersection alone relative to the other two sets. The reasons for this are: 1) that the WDI itself is the smallest data source in size and 2) that the corresponding query set comprises approximately 50 percent of the size of the original database. This value is about one to two orders of magnitude greater in comparison with the contents of the other two query sets. Because of this, we naturally expected the performance of the WDI to be better in the beginning, which, on the other hand, leaves less room for improvement later on. For the ZINC and PubChem queries, we were able to more than double the amount of queries that could be rebuilt exactly, and we could also significantly raise the amount of close analogues that were found. Therefore, the results we obtained show that we could achieve similar and very good performance for all the three query sets.

A case study is contained in Figure 3, which shows the best solutions obtained with the four fragment spaces for the nexavar (Sorafenib) query from Scheme 1. In this particular case, the query could only be exactly reconstructed with the BRICS_20k fragment space, which is reflected in the value of the corresponding MDL 'public' key. The complete list of hits can be found in the Supporting Information.

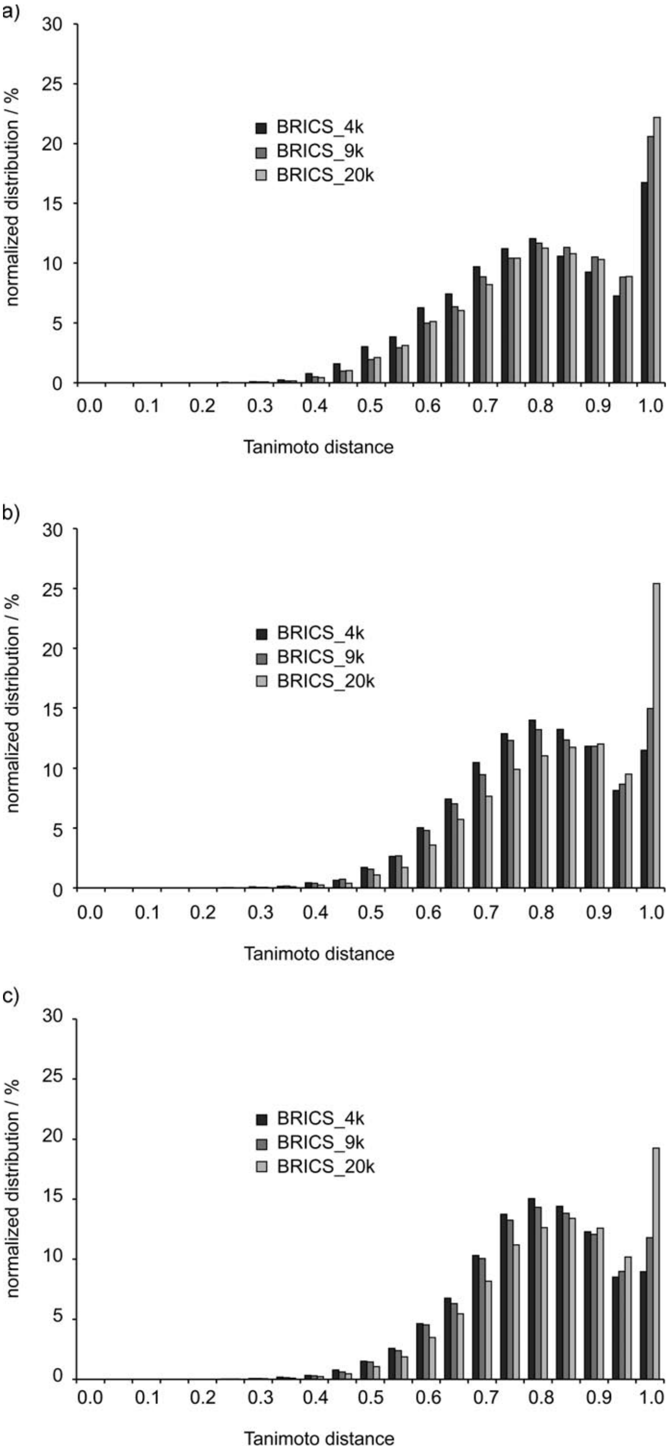


Figure 2. Comparisons of the performance of the enriched BRICS fragment spaces for the various query sets (a) WDI, b) ZINC, and c) PubChem). The histograms show the normalized distribution of the similarity values (Tanimoto distance of the respective MDL public keys) of the closest solution that could be constructed out of the corresponding fragment set with respect to each query molecule. See text for details.

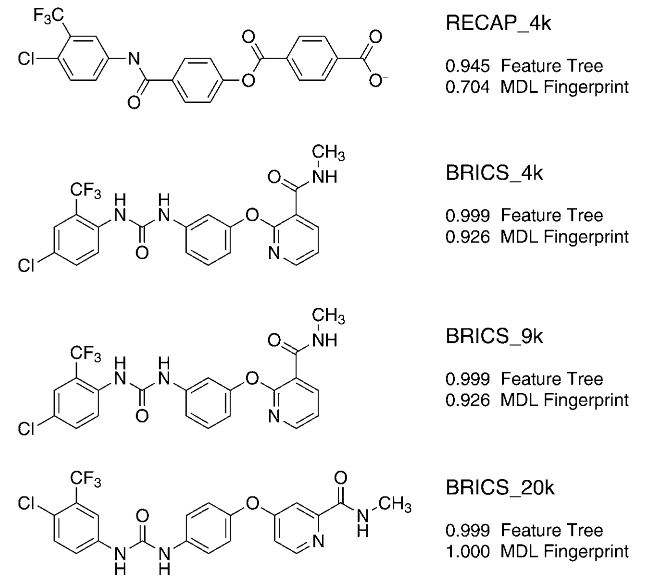


Figure 3. Case study showing the best re-ranked solutions obtained with the four fragment spaces for the nexavar (Sorafenib) query from Scheme 1. The corresponding similarity values are given next to each individual result.

The study presented herein was carried out to assess and improve the performance of current methods for automated retrosynthetic fragmentation and generation of molecules to support the drug-development process. It exploits the chemical information that is contained in existing compounds and combines it with knowledge about the principal feasibility of certain chemical motifs. We implemented a more elaborate and comprehensive model that better reflects medicinal chemistry concepts and is able to generate first candidates that are synthetically accessible. Using the new BRICS model, we generated three publicly available fragment spaces that can serve as a basis for various molecular design objectives, such as similarity searching, library design, and descriptor- or structure-based de novo design.[1]

The BRICS_4k intersection alone consists only of 'drug-like' fragments and performs best for obtaining molecules that have high similarity to compounds contained in the WDI. The enriched intersections BRICS_9k and BRICS_20k additionally contain varying amounts of fragments that were derived from compounds of vendor catalogues only, but still have a reasonable similarity to 'drug-like' fragments obtained from the WDI. The enriched intersections raise the performance such that a significant quantity of molecules can also be generated with a high degree of similarity to publicly available chemical matter.

Computational approaches for the generation of molecules are often believed to suggest structures that cannot be easily made experimentally. The concept presented herein considers synthetic concepts right from the beginning and in a more elaborate way. It also preserves promising chemical motifs contained in existing compounds. Both aspects significantly raise the quality of the results with respect to the similarity between the generated molecules and existing inhibitors or compounds from vendor catalogues.

We are currently further investigating various types of fragment collections with respect to their nature and properties. In this context, we want to derive a measure for balancing different properties of such a collection, with particular emphasis on structural complexity, the generality, and 'drug-likeness' of the underlying fragments. We think that this will take us another step forward in the direction of optimizing (fragment) libraries for specific applications and that it will also help us to generate better and more reliable predictions.

Acknowledgements

The authors are grateful to all partners of the NovoBench project and to the German Federal Ministry of Education and Research for partial funding (grant 313324A).

Keywords: chemical fragment spaces · computer-aided molecular design · fragment-based lead discovery · retrosynthesis · shredding

[1] *BRICS*, http://www.zbh.uni-hamburg.de/BRICS.
[2] P. J. Hajduk, J. Greer, *Nat. Rev. Drug Discovery* **2007**, *6*, 211.
[3] X. Q. Lewell, D. B. Judd, S. P. Watson, M. M. Hann, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511.
[4] H. Mauser, M. Stahl, *J Chem. Inf. Model.* **2007**, *47*, 318.
[5] *Daylight Theory Manual*, Daylight Chemical Information Systems, Inc. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.
[6] P. Maass, T. Schulz-Gasch, M. Stahl, M. Rarey, *J Chem. Inf. Model.* **2007**, *47*, 390.
[7] *World Drug Index*, Version 2004, Thomson, Philadelphia, PA (USA), 2004.
[8] J. J. Irwin, B. K. Shoichet, *J Chem. Inf. Model.* **2005**, *45*, 177.
[9] *ZINC 'drug-like' subset*, http://zinc.docking.org/subset1/3/index.html (version as of May 2, 2006).
[10] *Pipeline Pilot*, SciTegic, Inc., San Diego, CA (USA), **2007**.
[11] *Corina*, Version 3.4, Molecular Networks GmbH, Erlangen (Germany), **2007**.
[12] D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.
[13] *PubChem*, ftp://ftp.ncbi.nih.gov/pubchem/ (version as of November 10, 2007).
[14] M. Rarey, M. Stahl, *J. Comput. Aided Mol. Des.* **2001**, *15*, 497.
[15] M. Rarey, J. S. Dixon, *J. Comput. Aided Mol. Des.* **1998**, *12*, 471.
[16] *FTrees*, Version 1.5.5, BioSolveIT GmbH, Sankt Augustin (Germany), **2007**.
[17] J. L. Durant, B. A. Leland, D. R. Henry, J. G. Nourse, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273.
[18] T. T. Tanimoto, IBM Internal Report, November 17, 1957.
[19] R. J. Flower, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379.

Received: June 10, 2008
Revised: July 17, 2008
Published online on September 15, 2008