Similarity Searching and Scaffold Hopping in Synthetically Accessible Combinatorial Chemistry Spaces

Markus Boehm,*[,†] Tong-Ying Wu,[‡] Holger Claussen,[§] and Christian Lemmen[§]

Pfizer Global Research and Development, Eastern Point Road, Groton, Connecticut 06340, University of North Carolina, Chapel Hill, North Carolina 27599, and BioSolveIT GmbH, An der Ziegelei 75, D-53757 Sankt Augustin, Germany

Received June 28, 2007

Large collections of combinatorial libraries are an integral element in today's pharmaceutical industry. It is of great interest to perform similarity searches against all virtual compounds that are synthetically accessible by any such library. Here we describe the successful application of a new software tool CoLibri on 358 combinatorial libraries based on validated reaction protocols to create a single chemistry space containing over 10^{12} possible products. Similarity searching with FTrees-FS allows the systematic exploration of this space without the need to enumerate all product structures. The search result is a set of virtual hits which are synthetically accessible by one or more of the existing reaction protocols. Grouping these virtual hits by their synthetic protocols allows the rapid design and synthesis of multiple follow-up libraries. Such library ideas support hit-to-lead design efforts for tasks like follow-up from high-throughput screening hits or scaffold hopping from one hit to another attractive series.

Introduction

In the pharmaceutical industry, high-throughput screening (HTS[a]) of compound collections is frequently applied at the beginning of a drug discovery program to detect novel hits.[1–3] However, in many cases, HTS either fails to deliver any promising hit or the identified hits cannot be turned into lead compounds with desirable in vitro potency and selectivity. Furthermore, only a fraction of the initial leads survive in the further progress toward a candidate that makes it into clinical development.[4–6]

To address this problem of attrition, Pfizer has embarked in file enrichment (FE) programs in the past years to increase the quality of its corporate compound collection.[7–10] Enriching the collection was realized through investing in technologies that would make compound libraries larger, more chemically diverse, and more druglike. Well-designed libraries should produce higher quality compounds that will lead to an increase in the overall hit rate. The ultimate goal is to reduce the attrition rate for the resulting drug leads, thus increasing the overall efficiency and productivity further downstream in the discovery program.[6,11]

It has been estimated that the number of compounds that span the biologically relevant chemical space, that is, chemical compounds used by biological systems, exceeds 10^{60} molecules.[12–15] The strategy of FE to explore this vast chemical space efficiently has been to synthesize hundreds to a few thousand analogues around chemotypes giving the highest probability of covering the space and finding potential drug leads. To ensure druglike quality of FE compounds, molecules were designed to be rule-of-five compliant.[16] Chemotypes were selected focusing on maximum diversity, also by including structures with novel (and proprietary) scaffolds, to a more target-specific focus.[17–19] In the latter case, the increasing amount of genomic, proteomic, and structural data about druggable targets across gene families has provided information about structurally privileged scaffolds that can be used for the design of target-associated focused chemical libraries.[20–22]

Besides screening the newly augmented FE compound collection in an HTS campaign, it is also of great interest to access and search in silico the full virtual FE library. This virtual library is the collection of all possible compounds that can be combinatorially enumerated by FE synthetic protocols and is by several orders of magnitude larger than the real (synthesized) FE collection. Virtual screening has become an established in silico tool in drug discovery and is routinely used to identify potential hits as starting points for lead identification in discovery programs.[23–28] Molecular similarity searching is the method of choice when no information about the three-dimensional structure of the target is available.[29–33] Traditionally, similarity searching is performed by using one or multiple query structures and searching against a database of existing (or virtual) molecules. A broad variety of molecular descriptors can be used; the most prominent examples are structural keys[34,35] and molecule fingerprints[36–38] indicating the presence or absence of certain fragments or paths in the molecular graph, respectively. The drawback to these methods is that each molecule to be searched in the database has to exist as an explicit entity. Usually, this is not a problem for searching corporate compound collections or small combinatorial libraries which typically range in the order of 10^5-10^8 molecules. But large combinatorial libraries, especially collections thereof (such as the virtual FE library mentioned above), easily exceed this range by several orders of magnitude. Not only is it unfeasible to perform similarity searches for more than 10^8 virtual compounds, but it is often simply impractical to enumerate and store such large numbers of molecules. Consequently, there has been an increasing interest in computational methods recently to process large virtual combinatorial libraries in different ways.[39,40]

The topomer concept was introduced and applied by Cramer et al. to search virtual combinatorial libraries of 10^{13} molecules based on seven generalized libraries as part of Tripos' Chem-

* Corresponding author. Phone: +1-860-686-2031. E-mail: markus.boehm1@pfizer.com.
† Pfizer Global Research and Development.
‡ University of North Carolina.
§ BioSolveIT GmbH.

[a] Abbreviations: HTS, high-throughput screening; FE, file enrichment; CoLibri, Compound Library Toolkit; FTrees, Feature Trees; FTrees-FS, Feature Trees Fragment Spaces; WDI, World Drug Index; DY, Daylight; PP, Pipeline Pilot; USMILES, Unique SMILES.

10.1021/jm0707727 CCC: $40.75 © 2008 American Chemical Society
Published on Web 04/02/2008

Space platform.[41–43] Topomers are topologies of fragments from which a virtual compound can be generated. The comparison of topomers is achieved by examining the steric field properties of their fragment conformations. During the search, a query structure is split into fragments and the calculated steric fields of each fragment are compared with precalculated fields for fragments in the topomer library. Combining the similarity of each individual fragment gives the overall similarity of the virtual product molecules under consideration. The topomer shape similarity searching approach was shown to be effective for scaffold hopping in a number of cases considering the low Tanimoto similarities of their corresponding fingerprints. There are several notable differences between the topomer approach and the approach presented here (FTrees-FS). First, topomer search is a 3D method requiring the generation of 3D conformations of query and fragments as opposed to the purely topological fingerprint used by FTrees-FS. Second, topomer search is shape-based using the steric field of molecules, while the method we describe focuses on matching similar physicochemical features in a topologically plausible way. Third, with a topomer search the query molecule is fragmented and the search is based on fragment-to-fragment similarities, whereas FTrees-FS compares the entire query molecule to the fragments in the chemistry space for finding the best possible feature matches. Finally, while the topomer approach is based on few generalized reactions, the approach presented here constructs a chemistry space based on several hundred specific reaction protocols of combinatorial libraries.

Nikitin et al. constructed a large virtual diversity space of 10^{13} compounds derived from 400 combinatorial libraries described in the literature.[44] The original libraries were enriched by adding larger collections of chemical reagents from vendor catalogs. Their structure-based de novo design program generates candidate ligands within the time frame of several months. In comparison, the ligand-based similarity searching approach presented here requires only minutes to search.

Markush structure representations have been used by Barnard et al. to analyze and cluster virtual combinatorial libraries.[45] The internal Markush structure of a library is organized in a logic tree, in which the nodes store the fragments of the library and the edges indicate the logical and positional ways in which the fragments are connected to provide the products. Analysis of such Markush structures allows a more rapid generation of descriptors, physicochemical properties, and fragment-based fingerprints for the library products without enumerating them. This very compact representation allows for rapid enumeration and substructure searching. However, to the best of our knowledge it is not suitable for similarity searching as presented here.

FTrees-FS. New descriptors and comparison algorithms have been realized in the similarity searching program Feature Trees (FTrees)[46] and its extension Feature Trees Fragment Spaces (FTrees-FS),[47] which is capable of handling and searching large virtual combinatorial libraries without ever explicitly enumerating all possible product structures.

Originally, the program FTrees[46,48] was developed to perform similarity searches of query molecules against a database of (real or virtual) compounds. For each molecule, a feature tree descriptor is derived that captures all fragments and functional groups of the molecular structure as nodes and how they are linked to each other, thus defining the rough molecular topology as a tree. Subsequently, a feature profile is computed for each node in the tree that describes the physicochemical properties of the respective fragment or functional group. Features such as estimated volume, number of ring closures, donor, acceptor, aromatic, and hydrophobic properties are calculated.

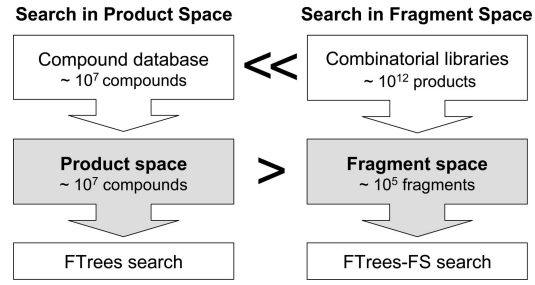

Figure 1. Comparison of similarity searches performed in product space vs fragment space. The number of enumerated combinatorial library products exceeds by far the number of compounds in corporate databases. However, the number of unique fragments is smaller than the number of molecules in the compound database. By searching in fragment space, FTrees-FS is able to cover efficiently the vast space of combinatorial chemistry space products.

The similarity between two molecules is calculated by comparing their associated feature trees by superposing (matching) similar subtrees onto each other. The advantage of using feature trees as descriptors for similarity searching is that they are capable of preserving the topology and physicochemical properties of the molecule. Since the underlying structural scaffold of the molecule is disregarded, some degree of fuzziness is introduced into the descriptor. This makes the method particularly suitable for areas such as lead or scaffold hopping.[49–58] The drawback to the FTrees method is—similar to the traditional search methods mentioned above—that only fully enumerated molecular structures can be considered during the search, which sets a limit to the number of compounds that can be stored and searched in a database.

As a consequence, FTrees-FS[47] was developed as an extension module to FTrees to perform similarity searches of large combinatorial chemistry spaces. In such a chemistry space the compounds represented by a combinatorial library are not stored as enumerated molecular structures (*product space*), but rather in the form of their building blocks and linkage rules (*fragment space*). The efficiency of searching combinatorial libraries encoded in their fragment space versus their product space can be easily explained by the different numbers of molecules that have to be compared during a similarity search. For instance, if a two-component combinatorial library with 1000 monomers (building blocks) each is searched in its enumerated form, 1000000 product structures need to be compared to a given query. In contrast, only 2000 monomers have to be considered in their corresponding fragment space. Extending this comparison to a large set of combinatorial libraries, the number of products can easily reach in excess of 10^{12} possible virtual products. Conducting the similarity search in the equivalent fragment space reduces the number of molecules to be searched to about 10^5 structures (Figure 1). This is even far less than the number of compounds in typical corporate collections. Furthermore, FTrees-FS does not simply compare monomers but rather generates products on-the-fly (based on a dynamic programming algorithm) that are similar to the query molecule. The particular nature of the feature tree descriptor allows the combination of all possible subtrees (representing the fragments) with sufficient similarity to successively larger subtrees and finally complete feature trees (representing the products) without ever enumerating all product structures. During the dynamic search procedure the selection and matching of subtrees is guided by locally maximizing the similarity to the query molecule.

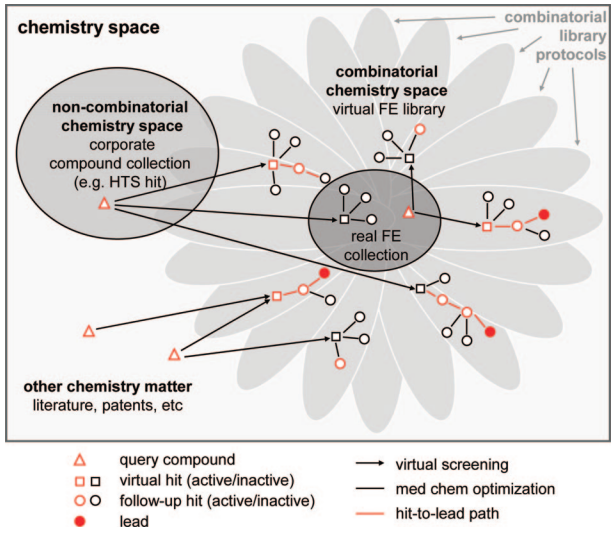

Figure 2. Searching in combinatorial chemistry space. The chemistry space of the virtual FE library contains a large number of combinatorial library protocols. Only a small portion of this space is covered by the real FE collection of synthesized compounds. Virtual screening with FTrees-FS allows us to search the much larger virtual FE library chemistry space.

So far, FTrees-FS similarity searching has not been applied to chemistry spaces derived from combinatorial chemistry libraries, but rather to more generally defined chemistry spaces.[47,59] These latter types of chemistry spaces usually consist of a few thousand fragments that can be combined based on a rule set that represents a limited number of basic chemical reactions. To build such a chemistry space, the retrosynthetic combinatorial analysis procedure (RECAP)[60] has been applied to drugs from the World Drug Index (WDI)[61] and other druglike molecule collections. Applying a set of retrosynthetic rules, the drug molecules are disconnected into their individual fragments, and together with their associated link types they are incorporated into the chemistry space.[47] This approach warrants a high degree of diversity of possible virtual products by allowing a large variety of fragments to be generated during the RECAP fragmentation process.[59] However, it comes at the expense of the a priori unknown synthetic feasibility of the discovered virtual hits since this type of chemistry space is not purely combinatorial in nature. Molecules generated from such a space contain a variable number of fragments and do not follow a unique synthetic reaction scheme.

CoLibri. In this paper, we describe the development and application of a new software tool CoLibri (Compound Library Toolkit),[62] which is capable of handling large numbers of combinatorial libraries, processing them into a database of fragments, and converting them into a single combinatorial chemistry space. Once such a fragment space is assembled it can be used for similarity searching with FTrees-FS. Figure 2 illustrates some of the most common virtual screening scenarios. Typically, search queries are HTS hits, published compounds from literature or patent applications, or hits from the real FE compound collection. The result of the similarity search is a set of virtual hits from the fragment space that are similar to the search query. Some of these initial hits might already exist as synthesized compounds, for example, as part of the FE collection, or they can likely be made in a straightforward manner using the synthetic protocols associated with the virtual hits. Validated hits showing biological activity can be further evaluated with additional hit follow-up libraries or optimized in a hit-to-lead medicinal chemistry campaign.

In the following text, we first present validation experiments to prove the feasibility of virtual screening in Pfizer's combinatorial chemistry space based on FE library protocols. Then we provide several application examples of similarity searches within this chemistry space. In Materials and Methods, we briefly highlight the strategy of computationally encoding combinatorial libraries and creating chemistry fragment spaces with CoLibri. Finally, in the Computational Appendix we describe in detail the generation of a chemistry fragment space using 358 combinatorial libraries based on FE reaction protocols. We also provide more information about the functionality of the CoLibri software.

Results and Discussion

The Pfizer chemistry fragment space consists of several hundred combinatorial libraries and over 10^{12} possible products. Within such a vast chemistry space it is difficult to assess the overall quality and correctness of the products dynamically generated during the FTrees-FS search. It is unfeasible to verify that all fragments from the various library protocols are connected properly according to a quite complex matrix of compatibilities. Nonetheless, our goal was to ensure that during the similarity search only valid product molecules are formed according to their associated library protocols. Only then can the identified virtual compounds be synthesized with high likelihood by applying the associated reaction conditions to the relevant building block monomers. Therefore, it was crucial for us to put the fragment space through a comprehensive validation before using it in real-life applications.

Validation of the Chemistry Fragment Space. Our approach to validate the accuracy of the chemistry fragment space was to generate a sample set of virtual product structures included in the combinatorial library collection. This was accomplished by randomly selecting five sets of monomers per protocol and connecting them according to their reaction scheme to form five virtual products for each of the 358 reaction protocols implemented in the chemistry space, resulting in 1790 virtual products in total. Each of these molecules successively served as query for FTrees-FS similarity searches in the fragment space. For each of the 1790 searches, the top 100 solutions with the highest similarities were retrieved, together with their respective rank and similarity score. A solution, which is a virtual product identified in the fragment space, was considered as a "hit" only if identical to the search query. We defined identity not only in the sense of representing the same chemical structure, but also if the product was annotated with the same reaction protocol and monomers used as the search query. In a best case scenario, for each of the five search queries of the 358 protocols such a hit would be identified as the top rank with a similarity score of 1.0. The validation outcome shows that for almost all (356 of 358) protocols at least three out of five times a hit was identified, which corresponds to a retrieval success rate of 99%.

Next, we inspected the rank distribution of the detected hits among the top 100 solutions captured for each search query. A total of 1210 out of 1418 hits were found among the first ten ranks, and another 87 hits could be retrieved between ranks 11 and 20. This provides an overall retrieval rate of 91% among the top fraction of the list, demonstrating that the similarity searches were able to identify the majority of hits at low rank numbers.

Furthermore, we were interested in determining to what extent similarity scores between identical search queries and hits deviate from their ideal value of 1.0. Such deviations can occur

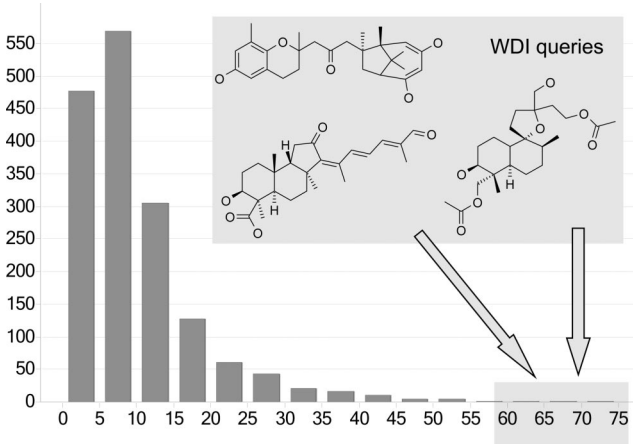

Figure 3. Distribution of compute times across the 1661 WDI query searches. For the majority of queries (90%), the search time was under 20 min while using a single processor machine. For only a few, more complex, query structures the search time tended to be longer (>60 min).

due to possible mismatches of atom or bond types between the feature trees of query and hit. For example, the bond type of an amide group is expected to be "amide", whereas the relevant acid and amine bonds in the fragment space are defined as "single". During the similarity search the amide bond type of a given amide search query and the generated solution will not be identical, thus resulting in a slight decrease of the similarity value from the ideal score. The distribution of similarity scores of the hits found in the validation showed that 1375 out of 1418 hits (97%) in the first ten ranks exhibit similarity scores of 0.95 or above. This verifies that the majority of hits come very close to their ideal similarity score value of 1.0.

Application of the Chemistry Fragment Space. After we successfully demonstrated that the chemistry fragment space is capable of generating valid product structures from their associated combinatorial library protocols, we embarked on an application to assess the coverage of druglike chemical space represented by this fragment space. We approached this task by selecting a broad range of known drugs from the WDI and tested if similar molecules could be found in the fragment space. First, the WDI database (version 2004)[61] was filtered by applying a few simple rules: organic filter (only molecules containing H, C, N, O, P, S, F, Cl, Br, and I atom types), rule-of-5 compliant, number of rings 1–4, maximum ring size 8, number of rotatable bonds less than 9. This was followed by a subset selection based on maximum diversity using Pipeline Pilot[63] FCFP_4 fingerprints which resulted in a representative subset of 1661 compounds. Similar to the validation procedure above, each of these WDI molecules served as query for FTrees-FS similarity searches. This time, only the top-ranked solution with highest similarity to the input query was taken into consideration, together with its similarity score and associated reaction protocol.

In order to be of practical relevance, a computational search procedure must have reasonable response times. Figure 3 shows the distribution of compute times across the 1661 searches. For the majority of queries (90%), the search time was below 20 min while using only a single processor machine (Intel Xeon 3.4 GHz, Red Hat Linux Enterprise). For 65% of the queries, the similarity search took even less than 10 min. This is a remarkable result, considering the size of the fragment search space comprises about 10^{12} theoretically possible product molecules. Note that a hypothetical search in the corresponding product space would take more than a month, assuming that it is possible to carry out each individual similarity calculation in just a microsecond. For certain WDI queries, the search time was significantly longer (>60 min). Such query structures (see Figure 3) were typically larger in size, less druglike, and in general showed a higher degree of complexity (high molecular weight, many functional groups, large number of rotatable bonds, complex ring systems). This is directly associated with an increase in compute time for such query molecules and can be explained by the underlying dynamic search algorithm of FTrees-FS. For each similarity search a lookup table needs to be dynamically generated to compare the query feature tree and all feature tree fragments from the fragment space. Since this lookup table represents the central data structure of the search algorithm, a larger similarity table generated by more complex query molecules directly leads to longer search times.

In the following step, the diversity of the synthetic chemistries identified during the similarity searches was analyzed. We wanted to make sure that the solutions generated from the FTrees-FS search algorithm covered a reasonable distribution of different reaction protocols. For example, search results that were largely dominated by only a few simple reactions such as amide bond formation with acids and amines would not be satisfying to a user who is interested in applying a broad range of potentially suitable reaction protocols to arrive at similar compounds. The frequency distribution of different reaction protocols retrieved for the solutions across the 1661 searches showed that about half (173 of 358; 48%) of the reaction protocols included in the fragment space were employed at least once to generate the top-ranked molecule. The most frequently occurring reaction protocols were as follows: reductive aminations; amide and sulfonamide formations; ether and ester formations; alkylation reactions with amines; nucleophilic aromatic substitution reactions; aryl−aryl (Suzuki) couplings; and various heteroaromatic ring formation reactions. Interestingly, the most frequently appearing protocol by far found in 151 searches was the reductive amination reaction. We want to emphasize that for the 358 protocols included in the fragment space there is inherently a certain degree of overlap and similarity in terms of their underlying synthetic chemistries, which explains why not all reaction protocols were employed. Also, we focused only on the top-ranked solutions; in a more realistic application scenario, hundreds or thousands of top-ranking solutions from a similarity search would be generated and analyzed for different reaction protocols among those products.

Next, in order to assess the coverage of druglike chemical space, we tested how similar the solutions were compared to their respective WDI queries. In Figure 4, the histogram shows the distribution of similarity scores of the 1661 queries. (Note that due to the nature of the feature tree descriptor the similarity scores from FTrees-FS are generally higher than similarity values derived from more classical descriptors like Daylight or Pipeline Pilot fingerprints. Therefore, the scores should not be compared directly to each other.) Lower similarity scores (<0.9) indicate that no product close in chemical structure could be generated from the fragment space. Products with scores in the medium range (0.9–0.95) usually exhibit a high degree of similarity in certain parts of the query. High similarity scores (>0.95) indicate that the search was able to find an analogue to the query, differing only in some minor features and topologies. At the extreme, even structures identical to the search query have been reproduced. After visual inspection of many generated structures at various levels of similarity we considered

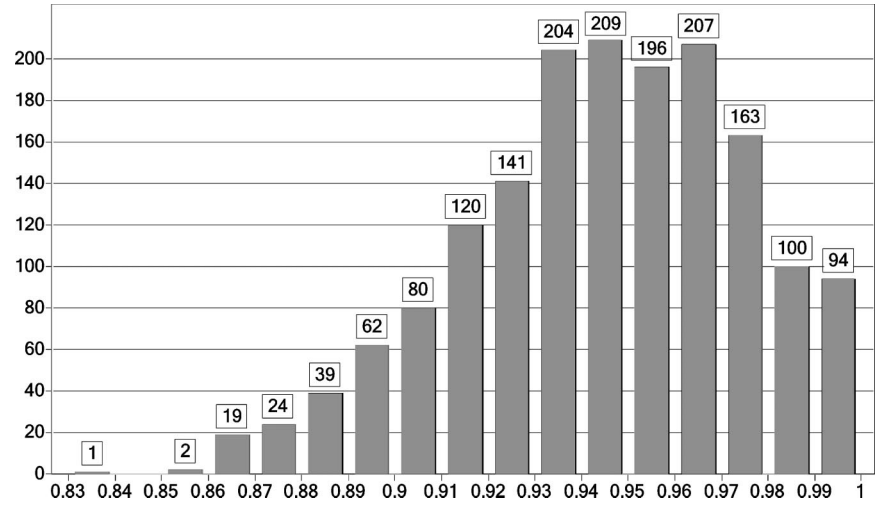

Figure 4. Similarity score distribution of the top ranking hits found for the WDI queries. 91% of the hits (1514 of 1661) showed a similarity score of >0.9, indicating that in most cases moderately to highly similar products compared to the query structures were identified.

the similarity score of 0.9 as a cutoff for finding a compound with evident similarity to a given query structure. From the 1661 WDI queries performed, 1514 structures showed a similarity score of >0.9, which corresponds to a coverage rate of 91%. This respectable and promising result led us to the conclusion that the generated fragment space covers a broad range of druglike chemical space and is capable of generating new compounds and compound series relevant for drug design.

Various examples of query molecules and their corresponding closest product structure generated from the fragment space grouped by similarity level (low, medium, high) are displayed in Figure 5. In the first group of cases with lower similarity (<0.9), the query structures were less druglike in general (Figure 5a). Many of them contained complex ring systems which are not necessarily expected in the more druglike fragment space, therefore making it almost impossible for the similarity search program to generate any closely related product structures. For instance, compound **1** contains a complex dibenzoquinolizinium system with four rings fused together.[64] The closest product found in the fragment space (similarity score 0.85) is the quinolinyl-aminoquinazoline derivative **2**.[65,66] The four-ring system of **1** is split into two separate two-ring systems with an amino group as a linker, replacing the quaternary amine from the query. In another example, the similarity search of search query **3** resulted in compound **4** (similarity score 0.87) where the tricyclic dihydrophenazine scaffold of the query is replaced by an indole ring connected to an aminobenzyl substituent.[67–69] In addition, only the methyl ester portion of the carboxylic acid methyl ester group of **3** was retained in **4**.

The next group of examples with solutions in the medium similarity range (0.9–0.95) contains products that are significantly closer to the query, often with structural deviations only in parts of the molecule (Figure 5b). Within this category there is the highest likelihood of finding potential lead-hopping candidates. For instance, the similarity search of **5** revealed the replacement of the central pyrroloindole scaffold by an indanyl piperazine ring in **6** (similarity score 0.918), thus preserving the basic amine present in both scaffolds.[70–72] Using search query **7** resulted in hit compound **8** (similarity score 0.935) where the phenothiazine heterocycle is replaced by a phenylindole scaffold.[73,74] The dimethylaminopropyl linker of **7**, which is important for activity, is fully retained in **8**. In another example, the central ketone group of **9** is substituted in **10** by an amide bond linker (similarity score 0.949).[75] Except for the two halogen substituents, the search was able to regenerate the identical quinazolinone heterocycle.

The last category of solutions with high similarity (>0.95) shows analogues that are very close compared to their queries (Figure 5c). As an example, the similarity search of **11** resulted in compounds that were almost identical to the query.[76] Among the top 500 solutions, the hit compounds **12**−**15** (similarity scores 0.95–0.98) were found, which are all known ligands binding to the same class of biogenic amine GPCRs.[77–80] The central carbamate group of query **11** was changed into a carboxylic ester or ether functionality. In the case of **15**, the diethylaminoethyl group of the query is replaced and the basic amine is bridged by a piperidine group to the rest of the molecule.

Interestingly, during the searches we also found numerous cases where the identical structure could be regenerated from the fragment space. For instance, the identical matching structure of compound **16** (Figure 6) was retrieved from the fragment space with a similarity score of 0.999 (see above for an explanation of slight deviations).[81] Using this last example, we further investigated whether our approach could provide quick and easy access to different synthetic chemistries based on a solution list for a given query. To this end, we retrieved a number of top-ranking products and checked their associated library protocols and their underlying synthetic reaction schemes. In particular, for the example WDI query **16** we analyzed the first 500 solutions and grouped them by their reaction protocols. Figure 6 shows the top-scoring molecule for each of the different reaction protocols. The top-ranked product **17** is structurally identical to the query, as mentioned before, and contains a sulfonylurea functional group formed during the synthetic reaction. On rank four, compound **18** is the first product from a different reaction protocol to synthesize sulfonamides.[82] The similarity to the sulfonylurea query **16** structure is preserved by including the amide functionality into the seven-membered lactam ring of **18**. Still very close to the query, on rank ten the first product **19** from the group of sulfonimides is retrieved.[83,84] The only difference between the query and this compound is the removal of one amino group next to the query cyclohexane ring, changing the sulfonylurea functional group into a sulfonimide analogue. Finally, another sulfonamide analogue **20** from a different reaction protocol is found on rank 50.[85] In this case, a methylene linker group was inserted between the sulfonamide and amide portion of **16**, transforming the sulfonylurea group

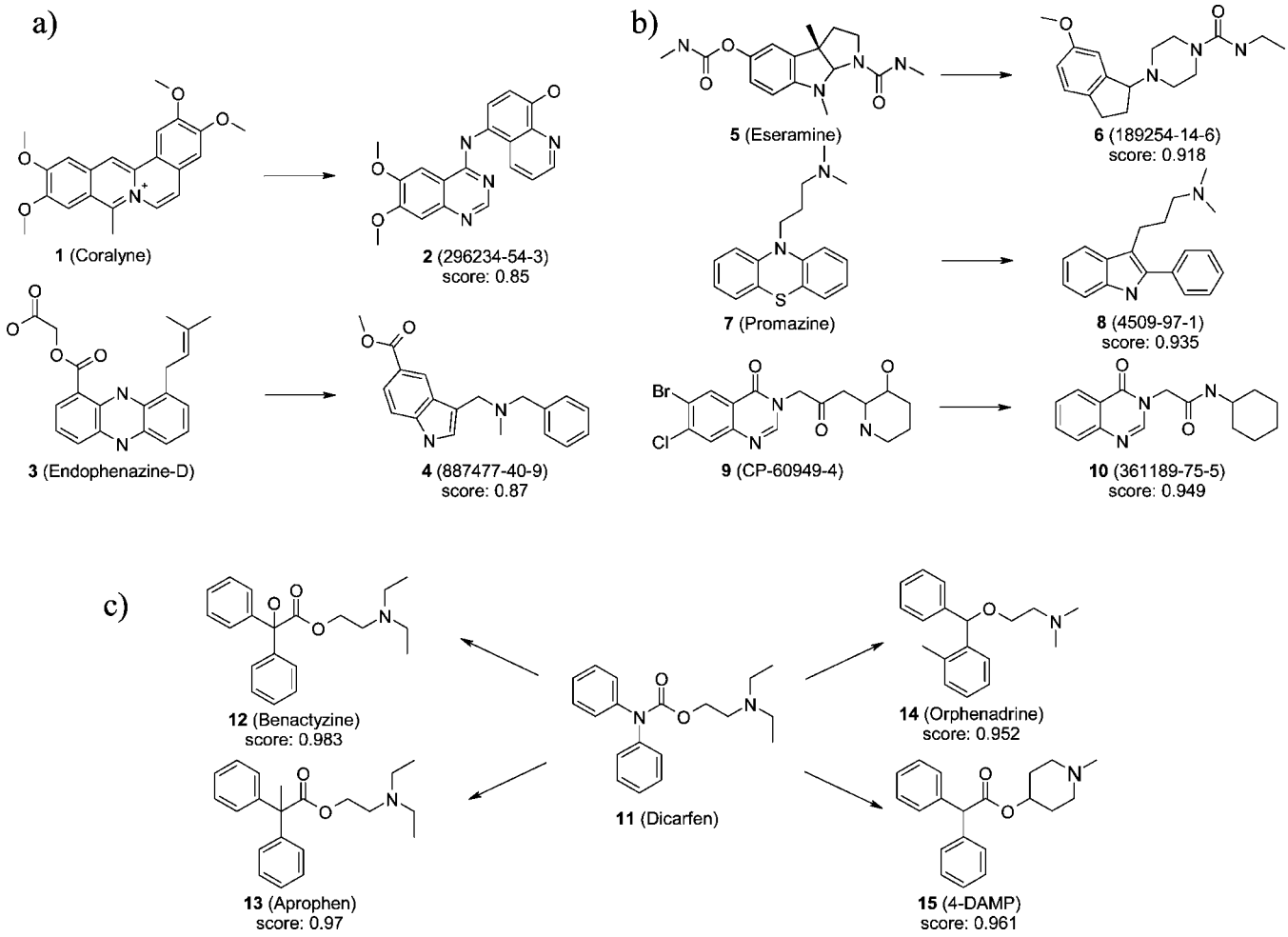

Figure 5. Examples of top-ranking solutions from the WDI queries found by the fragment space at (a) low (<0.9), (b) medium (0.9–0.95), and (c) high (>0.95) similarity levels. The WDI drug name or CAS registration number of the queries and obtained hits are shown together with the FTrees-FS similarity scores.

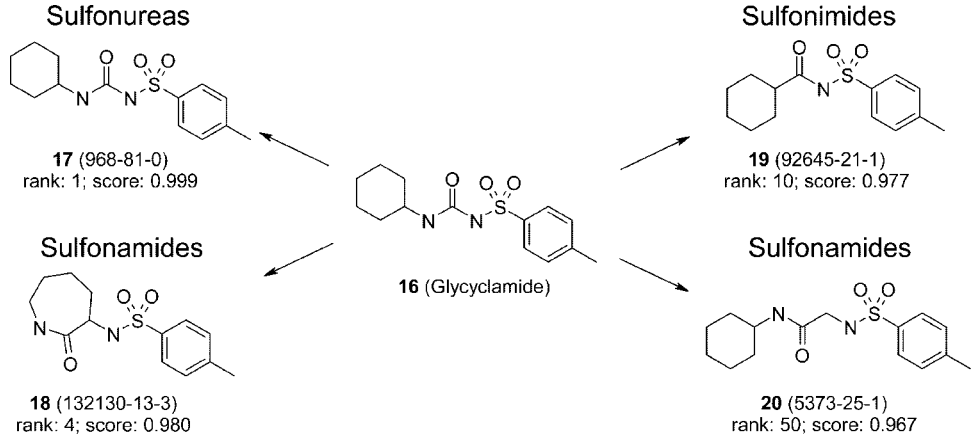

Figure 6. Examples of top-ranking solutions based on the WDI query **16**, highlighting the ability of the method to identify multiple synthetic chemistries to arrive at quite similar products. Different reaction protocols to synthesize sulfonylureas, sulfonimides, and sulfonamide products were retrieved. The CAS registration numbers of the obtained products are shown together with the FTrees-FS similarity scores and rank order in the solution list.

into a sulfonamide, both with a different chemistry protocol. This example nicely shows that by examining the various reaction protocols captured in the solution list it is possible to explore different synthetic chemistries for essentially the same target.

Scaffold Hopping in Target Classes. One of the key applications using the fragment space approach is to provide novel library design ideas for tasks such as HTS hit follow-up or scaffold hopping from a given active molecule (e.g., singleton compound, patent literature compound) to other attractive series which show activity for the same target. Therefore, it was of great importance to us to successfully demonstrate cases of scaffold hopping from one lead series of a given target class to another series of the same target family. We chose the serotonin

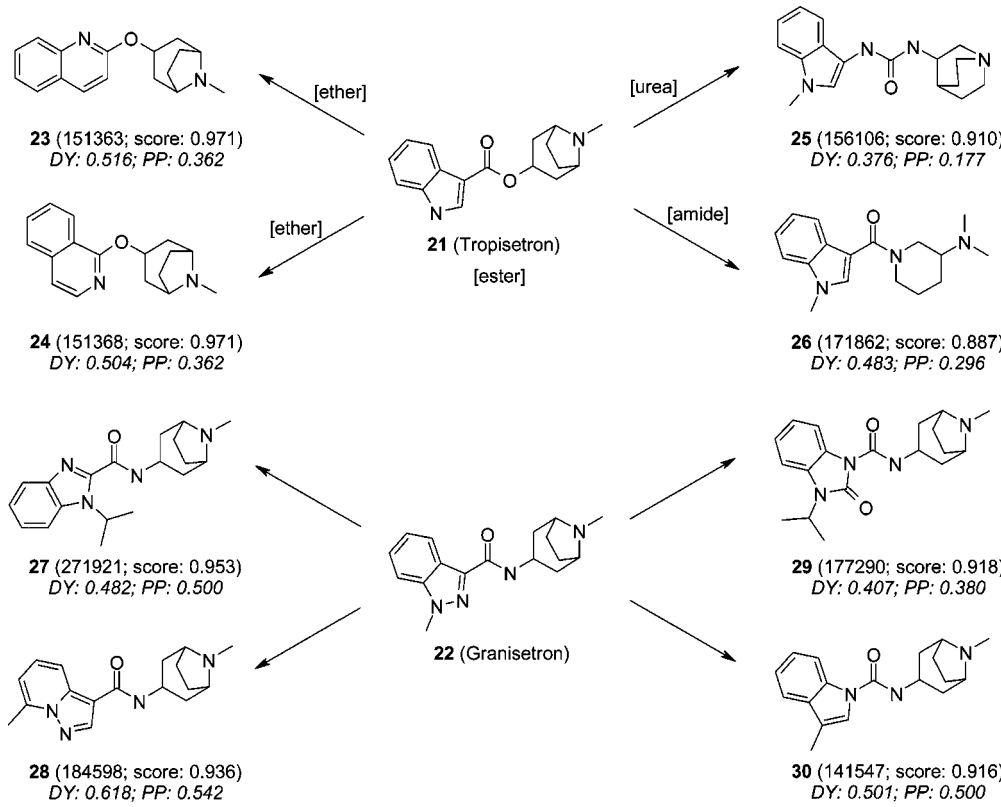

Figure 7. Examples of hits with known 5-HT3 antagonistic activity found in the fragment space. The two 5-HT3 antagonists **21** and **22** were used as search queries. The Prous Science Integrity registry numbers of the hits are listed together with their FTrees-FS similarity scores. For comparison, the Tanimoto similarities using Daylight (DY) and Pipeline Pilot (PP) FCFP_4 fingerprints are shown in italics. The various synthetic reaction protocols identified by **21** are indicated in parentheses.

5-HT3 receptor as an example target class. The 5-HT3 receptor subtype has been implicated in many brain functions. Since the stimulation by serotonin contributes to the sensation of nausea and vomiting, antagonists of this receptor exhibit antiemetic effects. For this study, a data set of 270 5-HT3 antagonists were extracted from the Prous Integrity database.[86] In addition, the two representative drugs **21** and **22**, which have been on the market since the early 1990s, were selected and used as search queries for FTrees-FS similarity searches in our fragment space.[87,88] To validate that other active compounds against the receptor could be identified, we examined the top 1000 solutions retrieved for each query to determine if any of them were identical to one of the 270 5-HT3 antagonists from the data set. Figure 7 shows the solutions from the FTrees-FS searches that were among the 270 known 5-HT3 antagonists. Aside from those eight products with identical structures to query molecules in the data set, many additional compounds were generated which showed close similarity to one of the known 5-HT3 antagonists. Naturally, some of the more interesting active hits are not published and documented as active in the literature. Unfortunately, for that reason we cannot disclose the structures of these hits.

The similarity scores of the solutions depicted in Figure 7 range from 0.887 to 0.971, indicating a medium to high similarity compared to the query. Remarkably, the scaffolds of all solutions as well as their underlying chemistries are quite distinct to those of the query structures, which underscores the strength of the feature tree descriptor in scaffold hopping. It is interesting to note that both queries—although relatively similar to each other—unveiled solutions that are rather different from each other. The first two hits **23** and **24** that are most similar to query **21** have a quinoline and isoquinoline ring, respectively, replacing the indole scaffold of **21**.[89] The query's ester functionality bridging the azabicyclic tropanyl ring system is substituted by an ether oxygen. Hence, in addition to modifying the scaffold a change in the underlying synthetic access has been identified (ester vs ether chemistry). The ester group of **21** has been replaced by an urea group in **25**, leading to another synthetic route (ester vs urea chemistry).[90] Furthermore, the bicyclic tropanyl ring of **21** was slightly altered to an azabicyclooctanyl group in **25**. In case of compound **26**, the tropanyl group of the query was replaced by a dimethyl amino piperidine substituent, at the same time transforming the ester linker to an amide (ester vs amide chemistry).[91] Using **22** as a query revealed distinct solutions with variations of the query's indazole scaffold. The structure of **27** shows a benzimidazole ring, whereas **28** contains a pyrrazolopyridine heterocycle instead.[92,93] In the case of solutions **29** and **30**, the indazole scaffold of query **22** is replaced by an oxobenzimidazole and indole ring, respectively.[94,95] At the same time, the synthetic access to the products compared to the query differs (amide to urea chemistry). The observation that most of the hits in Figure 7 do not replace the azabicyclic tropanyl ring in both queries is due to the fact that almost all of the 5-HT3 antagonists in the data set contain this very distinctive ring system. As mentioned before, other interesting solutions revealed by the similarity searches, including different scaffolds of the tropanyl ring, are not published in the literature and hence cannot be disclosed.

Next, we were interested to see if 2D similarity methods using established and routinely used descriptors such as Daylight[36] (DY) or Pipeline Pilot[63] (PP) FCFP_4 fingerprints are also in principle capable of retrieving the above identified hits with sufficiently high similarity. For such 2D fingerprints, it is generally accepted that molecules with Tanimoto similarity of

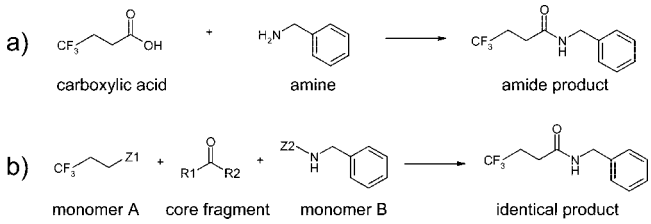

Figure 8. Original synthetic reaction scheme of a combinatorial library (a) is converted to a virtual reaction scheme (b) consisting of a core fragment and monomers A and B. The resulting product structures in both reaction schemes are identical.

0.7 and higher are considered similar to each other.[96,97] Using **21** as query, the Tanimoto similarity values of the four hits in Figure 7 ranged from 0.376 to 0.516 (DY) and from 0.177 to 0.362 (PP). The similarities of hits found for query **22** ranged from 0.407 to 0.618 (DY) and from 0.380 to 0.542 (PP). The overall average pairwise similarity between all eight hits was 0.438 (DY) and 0.342 (PP). All values are significantly below the similarity threshold of 0.7, indicating that the hits identified by the FTrees-FS searches would have not been found by standard 2D similarity searches using Daylight or Pipeline Pilot fingerprints. Also, we want to emphasize again that with these standard 2D similarity methods it is not possible to perform searches in a combinatorial chemistry space without exhaustive enumeration of the products, and searching within a 10^{12} product chemistry space would be practically impossible.

In summary, the above examples describing scaffold hopping in the same target class demonstrate that by examining the generated solutions from the FTrees-FS similarity searches it is possible to find potentially novel leads with distinct scaffolds that cannot be found by standard 2D similarity methods. Together with the reaction protocols associated with the identified hits, different synthetic chemistries can be explored allowing the user to select the most appropriate reaction protocols for a given task.

Materials and Methods

Encoding Combinatorial Libraries. Real synthetic combinatorial chemistry differs in many ways from its corresponding in silico representation. As an example, Figure 8 shows the synthetic reaction scheme for a combinatorial library to form amide products starting from carboxylic acids and primary or secondary amines as monomers. A conversion is necessary to turn this reaction into a fragment space, thereby making it accessible to FTrees-FS searches. The synthetic reaction scheme can be encoded as a virtual reaction by clipping the functional groups involved in the reaction from the monomers and replacing them with attachment points, also termed linker atoms or linkers. The linker atoms are identified by their labels, in the example shown as Z1 and Z2. At the same time, a so-called core fragment is created that connects the clipped monomers. In the exemplified reaction, a carbonyl group with two attachment points R1 and R2 represents the core fragment. The combinatorial library products are formed by connecting the R linker atoms from the core fragment and the Z linkers from the clipped monomers that are compatible to each other (i.e., R1 to Z1 and R2 to Z2). Two compatible linker atoms are connected by forming a bond between the respective neighbor atoms of the two linkers. The linker atoms themselves are eliminated as the bond is closed. In the case of the amide formation reaction, identical library products are enumerated by following the combination rules of the virtual reaction scheme (Figure 8).

This mechanism of describing the enumeration of a virtual reaction can be applied to any type of real combinatorial library. The synthetic reaction scheme is translated into a virtual reaction, which consists of one or more *core fragments* holding the R linker atoms, and a set of *clipped monomers* containing the Z-labeled linkers (Figure 9). A list of *compatibility rules* defines which linker atoms can be connected. In general, R and Z linkers with identical numbers of the same virtual reaction are compatible (R1 - Z1, R2 - Z2, . . ., R*n* - Z*n*). For the subsequent FTrees-FS searches of such libraries it is a prerequisite that all R and Z linker atoms involved in the virtual reaction are in a terminal position, i.e. they can only be connected to one neighboring atom. Furthermore, any ring closure, replacement, removal of protecting groups, and any other mechanism occurring in real synthetic reactions that cannot be directly translated into a virtual reaction scheme must be performed in a preprocessing step (see the Computational Appendix for more details).

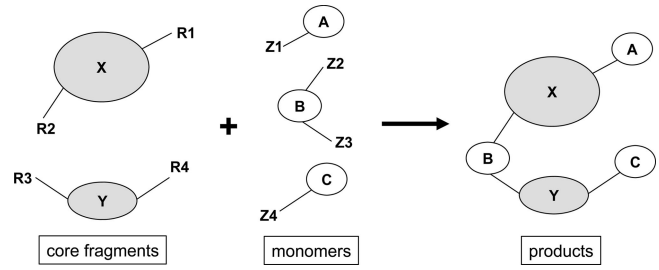

Figure 9. Every synthetic reaction can be translated into a virtual reaction scheme with core fragments (X, Y) and a set of clipped monomers (A, B, C). The products of the combinatorial library are formed by a set of compatibility rules defining which linker types can be connected to each other (R1 - Z1, R2 - Z2, etc).

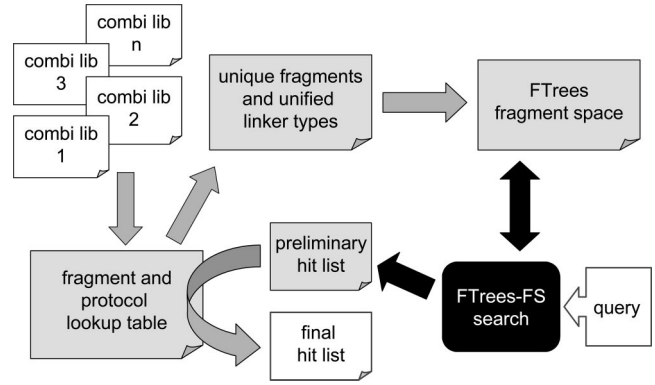

Figure 10. High-level overview of the workflow to generate a fragment space from a collection of combinatorial libraries and subsequent FTrees-FS searches. White boxes represent data input and output. Gray boxes and arrows indicate the data and processing steps of CoLibri. Black boxes and arrows relate to the actual FTrees-FS application. Note that the generation of a fragment space needs to be performed only once, while the searches based on such a fragment space can be continually performed. (See the Computational Appendix for more details.)

Generating Chemistry Fragment Spaces. To facilitate the encoding process of combinatorial libraries into chemistry fragment spaces, the toolkit CoLibri was developed. It allows the handling of large numbers of compounds or fragments and can be used to perform analyses and manipulations of the stored virtual molecules. It is also capable of storing and manipulating linker compatibility rules for valid fragment couplings. Chemistry spaces generated by CoLibri can be directly used by FTrees-FS. A high-level overview of the workflow for preparing a chemistry fragment space from a collection of combinatorial libraries, its connection to CoLibri, and the subsequent FTrees-FS searches is given in Figure 10.

We have applied CoLibri to build a chemistry fragment space using in-house FE library protocols. For this purpose a comprehensive collection of 358 validated, high-speed synthetic reaction protocols has been selected that are suitable for combinatorial chemistry libraries. A set of 138 two-, 202 three-, and 18 four-component reactions with a wide variety of reaction chemistries was chosen to maximize the structural diversity of the projected

chemistry space. Theoretically, if all 358 library protocols were fully enumerated it would result in more than 10^{12} possible virtual product structures. Clearly, this number exceeds the capability of any computational methodology operating in product space. In the Computational Appendix we illustrate in more detail how the in-house chemistry fragment space was generated. In addition, the functionality of the CoLibri software to compile large numbers of combinatorial libraries into manageable and searchable fragment spaces is provided.

Conclusions

A combinatorial chemistry space based on large numbers of validated FE library protocols was generated with the software tool CoLibri. This fragment space can be efficiently and exhaustively searched using the similarity search program FTrees-FS. The search results is a set of virtual compounds that are synthetically accessible by one or more of the existing library reaction protocols. Various validation experiments confirmed that the generated chemistry fragment space correctly defines trillions (10^{12}) of possible virtual product structures.

FTrees-FS was used to search within this fragment space based on a sizable number of query molecules taken from the WDI. The majority of these searches took less than 20 min on a single PC. The search results clearly indicate that the combinatorial chemistry space covers a broad range of druglike chemical space. In addition to generating new compounds relevant for drug design, it can also identify similar products with different synthetic chemistries. This allows the user to select the reaction protocol most suitable for a given task.

Finally, we demonstrated on the basis of a set of known 5-HT3 antagonists, using two marketed drugs for this target as queries, that it is possible to find novel leads with distinct scaffolds and different chemistries, thus allowing scaffold hopping from one compound to another attractive series active against the same target class.

Computational Appendix

Generation of In-House Chemistry Fragment Space. In a first step, each of the 358 protocols was translated into a virtual reaction scheme consisting of a unique core fragment and a list of reaction monomers in the form of clipped monomer fragments, as illustrated for the amide reaction example (see Figure 8). In general, two-component reactions always have one central core fragment, whereas three- and four-component reactions can consist of two or more core fragments. In case of the monomers, either two (A, B), three (A, B, C), or four (A, B, C, D) separate monomer lists are available for their respective two-, three-, and four-component reactions. The connecting reaction partners of the core fragments and monomers are labeled by their R and Z linker atoms, respectively, with matching numbers on both sides (R1 - Z1, R2 - Z2, etc.).

Next, the core fragments and monomer lists for each protocol were converted into SMILES[98] files. The originally R*n*- and Z*n*-labeled linker atoms were uniformly renamed by taking advantage of the SMILES mass specification [*m*]. In this context, the compatible linker pairs R1 and Z1 translate to [1*], R2 and Z2 to [2*], and so forth.

The following crucial step for generating a valid fragment space was to analyze the virtual reaction scheme to determine if any preprocessing steps are necessary. The two major instances that require preprocessing are ring-closure reactions and the occurrence of bridged (i.e., nonterminal) linker atoms, as described further below. Careful analysis of the 358 virtual reaction protocols allowed us to classify them into four groups: 146 protocols fulfilled the requirements of converting them into a fragment space without applying any preprocessing steps; 169 protocols revealed bridged linker atoms; 22 protocols were ring-closure reactions; and 21 protocols contained both bridged linker atoms and ring-closure formations. The latter three groups of reaction protocols all require preprocessing steps to make them compatible for the fragment space. Bridged linker atoms need to be shifted to terminal positions by applying certain clipping operations, and ring-closure reactions are partially enumerated to form the ring.

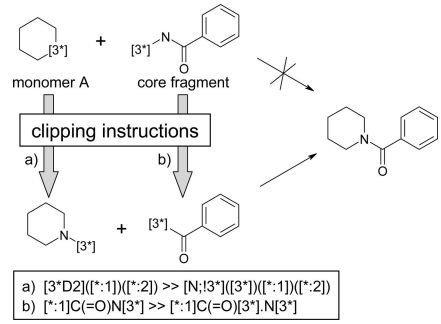

Figure 11. Example of preprocessing a reaction containing bridged linker atoms. Two necessary transformation steps are necessary to shift the nitrogen into the ring of monomer A and remove it from the core fragment: (a) identify linker [3*] with two anchor atoms of any type and replace it with a nitrogen atom with the now terminal linker [3*] attached to it while keeping the original two anchoring atoms untouched; (b) search for carboxamide substructure with a terminal linker [3*] and cleave the amide bond by inserting a dot between the carbonyl group and the nitrogen atom.

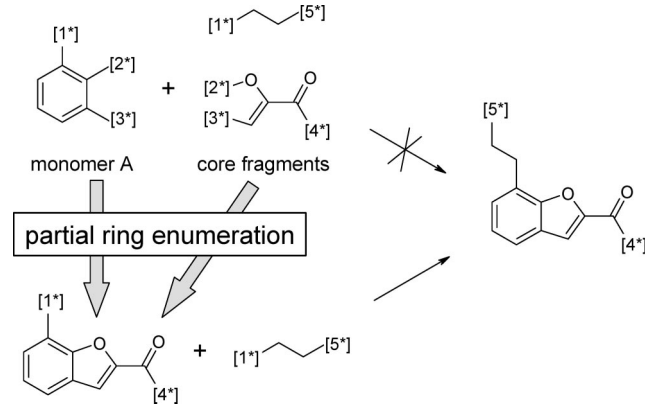

Figure 12. Example of preprocessing a reaction containing a ring formation as part of its reaction scheme. When the linker atoms [2*] and [3*] of monomer A and the corresponding core fragment are combined, they form a new ring. All variants of monomer A and the core fragment would need to be enumerated as part of the preprocessing step in order to generate all possible variants of the newly formed ring. Hence this step is called partial ring enumeration.

As an example of preprocessing, Figure 11 shows a reaction containing bridged linker atoms. Linker [3*] (mass label notation for R3) of monomer A is placed inside the six-membered ring. To move this bridged linker out of the ring into a terminal position the nitrogen from the core fragment can be shifted into the ring of monomer A. In CoLibri, this modification is accomplished by applying *clipping instructions*: the relevant substructure to be modified is captured by a SMARTS query, followed by a transformation of the original fragment via a SMIRKS expression.[36,99] In the exemplified reaction the two necessary transformation steps are to shift the nitrogen into the ring of monomer A and then remove it from the core fragment. The first clipping instruction (Figure 11a) identifies linker [3*] with two anchor atoms of any type (left side) and replaces it

with a nitrogen atom with the now terminal linker [3*] attached to it while keeping the original two anchoring atoms untouched (right side). The second expression (Figure 11b) searches for a carboxamide substructure with a terminal linker [3*] attached to the nitrogen (left side). The amide bond is cleaved by inserting a dot between the carbonyl group and the nitrogen atom, and an additional new linker [3*] is directly attached to the carbonyl group (right side). CoLibri later automatically removes the separated nitrogen atom formed as an intermediate byproduct during the clipping process.

Figure 12 illustrates an example protocol that contains a ring formation as part of its reaction scheme. Linker atoms [2*] and [3*] of monomers A (phenyl) and one of the core fragments (furan) are combined to form the benzofuran ring. Since ring closures cannot be handled by the FTrees-FS search algorithm, monomers A and the core fragment must be partially enumerated as part of the preprocessing routine. For this task, CoLibri provides functionality to select user-specified subsets of fragments and enumerates them according to the corresponding compatibility rules. The original fragments are then replaced by the partially enumerated fragments, together with a modified set of compatibility rules. In the example shown in Figure 12 the original fragments of monomer A are replaced by the partially enumerated benzofuran fragments with one linker atom [1*] left. During the FTrees-FS search, this linker will be combined with the remaining core fragment (ethyl group), following the compatibility rules of the modified virtual reaction scheme.

Protocols containing both a ring-closure step and bridged linker atoms have to be preprocessed twice. First, fragments forming the ring are identified and partially enumerated in order to obtain a new set of modified fragments. In a second step, bridged linker atoms of those modified fragments are processed as described above to shift them into a terminal position.

After modification of all reaction protocols requiring preprocessing, a single fragment lookup table was generated. During this step, all participating fragments from the 358 reaction protocols including their monomer and associated protocol names were collected. The fragment database consists of approximately 1.4 million fragments and serves as a lookup table during the postprocessing of search results (see below for details on postprocessing). Since this lookup table contains identical fragments from several protocols, CoLibri was used to remove any existing redundancy. All unique fragments represented in the entire generated fragment collection were identified, and new unified linkers required for nonredundant storage were created. This way, the original 1.4 million fragments were reduced to 89268 unique fragments and 7371 unified linker types. In a final step, the nonredundant fragment collection was converted into a fragment space file which serves as input to the FTrees-FS similarity search program. The Corina program[100,101] was used to convert the fragment collection from SMILES notation to a MOL2 file required by FTrees-FS. For this purpose, the "dummies" (-i dummies) and "mass to label" (-o m2l) options from Corina were used to copy the isotopic mass labels given in the SMILES input file into the corresponding atom name field in the MOL2 file. This way, for example, a linker type [1*] is converted to the "R1" linker atom. A modified Corina version was provided to us which is capable of handling isotopic mass labels with up to six digits instead of just three (this feature is now standard for Corina versions 3.4 and newer).[101] The total computing time of CoLibri covering all the steps to generate the final fragment space was about 12 h on a single processor (Intel Xeon 3.4 GHz, Red Hat Linux Enterprise). The majority of the compute time was used to identify and retrieve the unique fragments and to unify the linker types.

Functionality of CoLibri. The functionality of CoLibri that is necessary to compile a larger number of combinatorial libraries into a manageable and searchable fragment space consists of five components: (1) a mechanism to store fragments, identify duplicates, and select a unique representative; (2) functions to perform modifications to molecules; (3) a data structure to store and modify the compatibility of unique fragments; (4) import and export capabilities for different file formats; and (5) a postprocessing routine to relate back to the original input data.

Fragment Identification and Storage (1). CoLibri represents each fragment using a hashcode[102] for quick reference and a unique SMILES (USMILES) string[98] to capture the compound in a condensed form. The rationale for using both representations simultaneously is that the hashcode is better suited for storing and retrieving processes as it can be directly used as an address in a storage array. However, the hashcode cannot be used to regenerate the corresponding molecule. The USMILES, on the other hand, is a compressed representation of the molecule allowing storage of millions of molecules in memory, which can serve as input for numerous in silico screening operations. A fairly condensed representation without duplicates is required as the number of fragments to be represented can grow substantially. In our experiments, we found that in a redundant representation up to 25 times as many molecules would need to be considered. This is not surprising since monomers like amines, aldehydes or carboxylic acids are commonly used in numerous parallel reaction protocols. Currently, CoLibri is capable of handling up to a few hundred thousand fragments and keeping them in memory, which is essential for rapid duplicate identification and removal. CoLibri stores each input fragment in a lookup table connecting the unique representation together with the location of the original input. This reference is necessary to allow specific annotations to the fragment, for example, an in-house registry number or the associated reaction protocol.

Fragment Manipulation (2). As we outlined above, there are certain restrictions to the way fragments can be represented and combined in a virtual reaction scheme applicable to FTrees-FS searches. For example, virtual reactions containing fragments with bridged (i.e., nonterminal) linker atoms cannot be processed directly but require a preceding modification step. According to the compatibility rules explained before, every linker must be located in a terminal position with only one neighboring atom. Such cases are preprocessed by adding one atom from the side chain of the reaction partner to the fragment containing the bridged linker (usually a ring system) and simultaneously removing it from the donating side chain (see Figure 11). In order to facilitate these types of transformations in an automated fashion, CoLibri provides mechanisms for subgraph matching and replacement. The user specifies in a reaction transforming expression the substructure to be modified as a SMARTS query, followed by the SMIRKS pattern of the replacement group.[36,99]

Other examples that cannot be represented directly but instead must be handled in two consecutive steps are ring-closure reactions. In a preprocessing step, the two (or more) fragments that form the ring must be connected according to the corresponding compatibility rules (see Figure 12). If this involves only two individual fragments, then just one fragment containing the closed ring is created. More generally, there may be multiple variations in the fragments forming the ring. In these cases, a

partial enumeration is performed involving only the fragments participating in the ring formation according to their connection rules. This permits the generation of a new set of fragments that contain all the ring closures. Subsequently, the original ring-forming fragments as well as the connection rules defining the ring closure need to be removed. Finally, in the second step the newly created fragments can be combined as usual following the compatibility rules of the (modified) virtual reaction scheme of the combinatorial library.

Fragment Compatibility Handling (3). Since CoLibri keeps all fragments and their corresponding linker atoms in memory, a single compatibility matrix (R linkers vs Z linkers) is created in order to facilitate the book-keeping of all the combination rules implied in the different virtual reactions. As an example, the combinatorial library illustrated in Figure 9 consists of three components (monomers A, B, C) and two core fragments, resulting in eight linker types (R1 to R4 and Z1 to Z4). The simplistic rule, stating that linker atoms with identical numbers within the same reaction are compatible, translates to the few corresponding entries in the compatibility matrix. However, by adding more and more protocols the number of linkers increases. Since fragments occur in multiple protocols, more entries appear in certain cells of the matrix indicating such multiple compatibilities. Note that with two partially overlapping monomer lists A and B from two different protocols the monomers need to be treated as three subsets for nonredundant storage: monomers A without B (A\B), monomers B without A (B\A), and the intersection of monomers A and B (A ∩ B). While the first two monomer subsets may require only one linker type each, the latter overlapping subset requires a third linker type which is compatible with both core fragments of the respective two protocols. Due to this effect, the number of linkers may become quite large.

Import and Export of Fragment Spaces (4). The two major tools capable of searching fragment spaces, FTrees-FS[47] and FlexNovo,[103] utilize essentially the same syntax. It consists of a file with all fragments containing the linker atoms, and another file storing the compatibility matrix. One difference is that FlexNovo requires 3D conformations of the molecules, requiring the specification of bond lengths and angles for the combination of fragments. CoLibri is able to export fragment spaces in both flavors. In terms of input options, fragments can either be provided as SD, SMILES, or MOL2 files. As mentioned previously, CoLibri utilizes internally a USMILES representation. In the case of FlexNovo, 3D coordinates are required which can be calculated on-the-fly using Corina.[100,101]

Postprocessing of Output (5). In addition to the molecular structures of the resulting search hits, it is often desirable that this output can be mapped back to the original input data. This allows the user to retrieve information such as reaction protocol or in-house registry numbers associated with the output molecule and to embed this functionality seamlessly into existing workflows. For this purpose, as previously mentioned, CoLibri stores all original input data by reference in a lookup table. The name of a particular virtual product structure found in the fragment space is represented as concatenation of hashcodes of the combined fragments. By identifying the unique core fragment of the virtual product, which is linked to the reaction protocol, CoLibri can determine the original monomer names in the lookup table by mapping them to the hashcodes of the virtual product (see Figure 10). This way it is possible to identify the specific combinatorial library protocol and monomers that are required to form the particular virtual product structure at hand.

A further postprocessing step needs to be applied to the preliminary hit list to filter out all solutions that are not complete product structures because they have been early terminated during the fragment space search. The similarity search algorithm of FTrees-FS itself is not able to detect if a virtual product represents a fully enumerated structure of a particular combinatorial library protocol. It will only combine fragments according to the underlying compatibility rules of the fragment space, until a virtual product with the highest possible similarity against a given query molecule is found. By chance, this virtual product may be incomplete; for example, in a three-component reaction the core fragment may only combine with two of the three required monomer fragments, hence forming a partially enumerated product. In such cases, the apparently incomplete solution needs to be eliminated from the final output hit list. In FTrees-FS this is achieved by terminating all unsatisfied linkers in the virtual products with a potassium (K) atom, which is otherwise not occurring in druglike molecules. The preliminary list of virtual products can easily be filtered by CoLibri to exclude all solutions containing a potassium atom, since these represent only incomplete virtual product structures.

Acknowledgment. Numerous colleagues contributed to the project described in this paper. The authors are grateful to all of them for their support. In particular, M.B. thanks Alan Mathiowetz, Gregory Bakken, and David Piotrowski (Pfizer) for their support and many fruitful discussions. T.W. is grateful to Alexander Tropsha for his support. C.L. and H.C. gratefully acknowledge Ingo Dramburg and Markus Lilienthal (BioSolveIT) for their work on the implementation of CoLibri. Special thanks go to Christof Schwab (Molecular Networks) for providing us with the quick access to a modified version of the Corina program.

References

(1) Fox, S.; Farr-Jones, S.; Sopchak, L.; Boggs, A.; Comley, J. High-throughput screening: searching for higher productivity. *J. Biomol. Screen.* **2004**, *9*, 354–8.

(2) Posner, B. A. High-throughput screening-driven lead discovery: meeting the challenges of finding new therapeutics. *Curr. Opin. Drug Discov. Devel.* **2005**, *8*, 487–94.

(3) Hüser, J.; Lohrmann, E.; Kalthof, B.; Burkhardt, N.; Brüggemeier, U.; Bechem, M. High-throughput Screening for Targeted Lead Discovery. In *High-Throughput Screening in Drug Discovery*; Hüser, J., Ed.; Wiley–VCH: Weinheim, Germany, 2006; pp 15–36.

(4) Lahana, R. How many leads from HTS? *Drug Discov. Today* **1999**, *4*, 447–448.

(5) Ramesha, C. S. How many leads from HTS? - Comment. *Drug Discov. Today* **2000**, *5*, 43–44.

(6) Carnero, A. High throughput screening in drug discovery. *Clin. Transl. Oncol.* **2006**, *8*, 482–90.

(7) Ratner, M. L. File Enrichment: The Way Out of Pharma's Productivity Crisis? In *Start-Up*; Windhover Information, Inc: Norwalk, CT, 2002.

(8) Milne, G. M. Pharmaceutical productivity: the imperative for new paradigms. In *Annual Reports in Medicinal Chemistry*; Academic Press: New York, 2003; Vol. 38, Chapter 35, pp 383–396.

(9) Estep, K. File Enrichment and Hit Follow Up: Evolution and Examples. In *ALA LabFusion*; Boston, MA, 2004.

(10) Smith, G. F. Enabling HTS Hit Follow up via Chemoinformatics, File-Enrichment and Outsourcing. In *High Throughput Medicinal Chemistry II*; MMS Conferencing & Events Ltd., Institute of Physics: London, 2006.

(11) Borman, S. Improving Efficiency. To eliminate R&D bottlenecks, drug companies are evaluating all phases of discovery and development and are using novel approaches to speed them up. *Chem. Eng. News* **2006**, *84*, 56–78.

(12) Dobson, C. M. Chemical space and biology. *Nature* **2004**, *432*, 824–8.

(13) Lipinski, C.; Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **2004**, *432*, 855–61.

(14) 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*, 17272–7.

(15) Gorse, A. D. Diversity in medicinal chemistry space. *Curr. Top. Med. Chem.* **2006**, *6*, 3–18.

(16) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.

(17) Goodnow, R. A.; Guba, W.; Haap, W. Library design practices for success in lead generation with small molecule libraries. *Comb. Chem. High Throughput Screen.* **2003**, *6*, 649–60.

(18) Rose, S.; Stevens, A. Computational design strategies for combinatorial libraries. *Curr. Opin. Chem. Biol.* **2003**, *7*, 331–9.

(19) Young, S. S.; Ge, N. Design of diversity and focused combinatorial libraries in drug discovery. *Curr. Opin. Drug Discov. Devel.* **2004**, *7*, 318–24.

(20) DeSimone, R. W.; Currie, K. S.; Mitchell, S. A.; Darrow, J. W.; Pippin, D. A. Privileged structures: applications in drug discovery. *Comb. Chem. High Throughput Screen.* **2004**, *7*, 473–94.

(21) Krier, M.; Bret, G.; Rognan, D. Assessing the scaffold diversity of screening libraries. *J. Chem. Inf. Model.* **2006**, *46*, 512–24.

(22) Kubinyi, H. Chemogenomics in drug discovery. *Ernst Schering Res. Found. Workshop* **2006**, 1–19.

(23) Lengauer, T.; Lemmen, C.; Rarey, M.; Zimmermann, M. Novel technologies for virtual screening. *Drug Discov. Today* **2004**, *9*, 27–34.

(24) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862–5.

(25) Stahura, F. L.; Bajorath, J. Virtual screening methods that complement HTS. *Comb. Chem. High Throughput Screen.* **2004**, *7*, 259–69.

(26) Stoermer, M. J. Current status of virtual screening as analysed by target class. *Med. Chem.* **2006**, *2*, 89–112.

(27) Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov. Today* **2006**, *11*, 580–94.

(28) Kubinyi, H. Success Stories of Computer-Aided Design. In *Computer Applications in Pharmaceutical Research and Development,* Ekins, S., Ed.; John Wiley & Sons: New York, 2006; pp 377–424.

(29) Lemmen, C.; Lengauer, T. Computational methods for the structural alignment of molecules. *J. Comput. Aided Mol. Des.* **2000**, *14*, 215–32.

(30) Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204–18.

(31) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* **2006**, *11*, 1046–53.

(32) Downs, G. M.; Willett, P. Similarity Searching in Databases of Chemical Structures. In *Reviews in Computational Chemistry*; Kenny B. Lipkowitz, D. B. B., Ed.; Wiley-VCH: Hoboken, NJ, 2007; pp 1–66.

(33) Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov. Today* **2007**, *12*, 225–33.

(34) MDL. MDL Information Systems Inc., http://www.mdli.com/.

(35) McGregor, M. J.; Pallai, P. V. Clustering of large databases of compounds: using MDL "Keys" as structural descriptors. *J. Chem. Inf. Model.* **1997**, *37*, 443–8.

(36) Daylight. Daylight Chemical Information Systems Inc., http://www.daylight.com/.

(37) McGregor, M. J.; Muskal, S. M. Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. *J. Chem. Inf. Model.* **1999**, *39*, 569–74.

(38) Godden, J. W.; Xue, L.; Stahura, F. L.; Bajorath, J. Searching for molecules with similar biological activity: analysis by fingerprint profiling. *Pac. Symp. Biocomput.* **2000**, 566–75.

(39) Schneider, G.; Lee, M. L.; Stahl, M.; Schneider, P. De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. *J. Comput. Aided Mol. Des.* **2000**, *14*, 487–94.

(40) Langer, T.; Wolber, G. Virtual combinatorial chemistry and in silico screening: Efficient tools for lead structure discovery. *Pure Appl. Chem.* **2004**, *76*, 991–996.

(41) Cramer, R. D.; Patterson, D. E.; Clark, R. D.; Soltanshahi, F.; Lawless, M. S. Virtual Compound Libraries: A New Approach to Decision Making in Molecular Discovery Research. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1010–1023.

(42) Cramer, R. D.; Poss, M. A.; Hermsmeier, M. A.; Caulfield, T. J.; Kowala, M. C.; Valentine, M. T. Prospective identification of biologically active structures by topomer shape similarity searching. *J. Med. Chem.* **1999**, *42*, 3919–33.

(43) Andrews, K. M.; Cramer, R. D. Toward general methods of targeted library design: topomer shape similarity searching with diverse structures as queries. *J. Med. Chem.* **2000**, *43*, 1723–40.

(44) Nikitin, S.; Zaitseva, N.; Demina, O.; Solovieva, V.; Mazin, E.; Mikhalev, S.; Smolov, M.; Rubinov, A.; Vlasov, P.; Lepikhin, D.; Khachko, D.; Fokin, V.; Queen, C.; Zosimov, V. A very large diversity space of synthetically accessible compounds for use with drug design programs. *J. Comput. Aided Mol. Des.* **2005**, *19*, 47–63.

(45) Barnard, J. M.; Downs, G. M.; von Scholley-Pfab, A.; Brown, R. D. Use of Markush structure analysis techniques for descriptor generation and clustering of large combinatorial libraries. *J. Mol. Graph Model.* **2000**, *18*, 452–63.

(46) Rarey, M.; Dixon, J. S. Feature trees: a new molecular similarity measure based on tree matching. *J. Comput. Aided Mol. Des.* **1998**, *12*, 471–90.

(47) Rarey, M.; Stahl, M. Similarity searching in large combinatorial chemistry spaces. *J. Comput. Aided Mol. Des.* **2001**, *15*, 497–520.

(48) Rarey, M.; Hindle, S.; Maass, P.; Metz, G.; Rummey, C.; Zimmermann, M. Feature Trees: Theory and Applications from Large-scale Virtual Screening to Data Analysis. In *Pharmacophores and Pharmacophore Searches*, Thierry Langer, R. D. H., Ed.; Wiley–VCH: Weinheim, Germany, 2006; pp 81–116.

(49) Cramer, R. D.; Jilek, R. J.; Guessregen, S.; Clark, S. J.; Wendt, B.; Clark, R. D. "Lead hopping. Validation of topomer similarity as a superior predictor of similar biological activities. *J. Med. Chem.* **2004**, *47*, 6777–91.

(50) Good, A. C.; Hermsmeier, M. A.; Hindle, S. A. Measuring CAMD technique performance: a virtual screening case study in the design of validation experiments. *J. Comput. Aided Mol. Des.* **2004**, *18*, 529–36.

(51) Evers, A.; Hessler, G.; Matter, H.; Klabunde, T. Virtual screening of biogenic amine-binding G-protein coupled receptors: comparative evaluation of protein- and ligand-based virtual screening protocols. *J. Med. Chem.* **2005**, *48*, 5448–65.

(52) Hessler, G.; Zimmermann, M.; Matter, H.; Evers, A.; Naumann, T.; Lengauer, T.; Rarey, M. Multiple-ligand-based virtual screening: methods and applications of the MTree approach. *J. Med. Chem.* **2005**, *48*, 6575–84.

(53) Barker, E. J.; Buttar, D.; Cosgrove, D. A.; Gardiner, E. J.; Kitts, P.; Willett, P.; Gillet, V. J. Scaffold hopping using clique detection applied to reduced graphs. *J. Chem. Inf. Model.* **2006**, *46*, 503–11.

(54) Renner, S.; Schneider, G. Scaffold-hopping potential of ligand-based similarity concepts. *Chem. Med. Chem.* **2006**, *1*, 181–5.

(55) Stiefl, N.; Zaliani, A. A knowledge-based weighting approach to ligand-based virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 587–96.

(56) Zhang, Q.; Muegge, I. Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: ranking, voting, and consensus scoring. *J. Med. Chem.* **2006**, *49*, 1536–48.

(57) Bergmann, R.; Linusson, A.; Zamora, I. SHOP: Scaffold HOPping by GRID-Based Similarity Searches. J. Med. Chem. 2007.

(58) Zhao, H. Scaffold selection and scaffold hopping in lead generation: a medicinal chemistry perspective. *Drug Discov. Today* **2007**, *12*, 149–55.

(59) Mauser, H.; Stahl, M. Chemical Fragment Spaces for de novo Design. *J. Chem. Inf. Model.* **2007**, *47*, 318–24.

(60) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP–retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Model.* **1998**, *38*, 511–22.

(61) WDI. World Drug Index, http://scientific.thomson.com/products/wdi/.

(62) BioSolveIT. BioSolveIT GmbH, http://www.biosolveit.de/.

(63) SciTegic. SciTegic Inc., http://www.scitegic.com/.

(64) Zee-Cheng, K. Y.; Paull, K. D.; Cheng, C. C. Experimental antileukemic agents. Coralyne, analogs, and related compounds. *J. Med. Chem.* **1974**, *17*, 347–51.

(65) Uckun, F. M.; Liu, X.-p.; Narla, R. K. Preparation of quinazolines as antitumor agents. WO 2000056720, 2000; Parker Hughes Institute, St. Paul, MN.

(66) Yiv, S.; Li, M.; Uckun, F. M. Preparation of quinazolines for micellar pharmaceuticals for treatment of allergy and cancer. WO 2000056338, 2000; Parker Hughes Institute, St. Paul, MN.

(67) Gebhardt, K.; Schimana, J.; Krastel, P.; Dettner, K.; Rheinheimer, J.; Zeeck, A.; Fiedler, H. P. Endophenazines A−-D, new phenazine antibiotics from the arthropod associated endosymbiont Streptomyces anulatus. I. Taxonomy, fermentation, isolation and biological activities. *J. Antibiot.* **2002**, *55*, 794–800.

(68) Krastel, P.; Zeeck, A.; Gebhardt, K.; Fiedler, H. P.; Rheinheimer, J. Endophenazines A-D, new phenazine antibiotics from the athropod associated endosymbiont Streptomyces anulatus II. Structure elucidation. *J. Antibiot.* **2002**, *55*, 801–6.
(69) Lindquist, C.; Ersoy, O.; Somfai, P. Parallel synthesis of an indole-based library via an iterative Mannich reaction sequence. *Tetrahedron* **2006**, *62*, 3439–3445.
(70) Robinson, B. Alkaloids of Physostigma Venenosum. Iv. The Synthesis of Eseramine. *J. Chem. Soc.* **1965**, *33*, 3336–9.
(71) Yu, Q. S.; Atack, J. R.; Rapoport, S. I.; Brossi, A. Synthesis and anticholinesterase activity of (−)-N1-norphysostigmine, (−)-eseramine, and other N(1)-substituted analogues of (−)-physostigmine. *J. Med. Chem.* **1988**, *31*, 2297–300.
(72) Mattson, R. J.; Catt, J. D.; Keavy, D.; Sloan, C. P.; Epperson, J.; Gao, Q.; Hodges, D. B.; Iben, L.; Mahle, C. D.; Ryan, E.; Yocca, F. D. Indanyl piperazines as melatonergic MT2 selective agents. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 1199–1202.
(73) Julia, M.; Melamed, R.; Gombert, R. Research in the indole series. XVI. 2-Aryltryptamines and homologous amines. *Ann. Inst. Pasteur* **1965**, *109*, 343–62.
(74) Ganellin, C. R.; Hollyman, D. R.; Ridley, H. F. Aminoalkylation of metal derivatives of indole. II. Coupling of indolylmagnesium iodides with haloalkylamines. *J. Chem. Soc. C* **1967**, 2220–5.
(75) Waletzky, E.; Berkelhammer, G.; Kantor, S. Quinazolinones for treating coccidiosis. US 3320124, 1967; American Cyanamid Co.
(76) Sekera, A.; Hruby, J.; Vrba, C.; Lebduska, J. Anesthetics of carbamic acid series. *Chem. Listy Vedu Prum.* **1950**, *44*, 275–6.
(77) Robitscher, J. B.; Pulver, S. E. Orphenadrine in the treatment of depression; a preliminary study. *Am. J. Psychiatry* **1958**, *114*, 1113–5.
(78) Amitai, G.; Herz, J. M.; Bruckstein, R.; Luz-Chapman, S. The muscarinic antagonists aprophen and benactyzine are noncompetitive inhibitors of the nicotinic acetylcholine receptor. *Mol. Pharmacol.* **1987**, *32*, 678–85.
(79) Tumiatti, V.; Recanatini, M.; Minarini, A.; Melchiorre, C.; Chiarini, A.; Budriesi, R.; Bolognesi, M. L. Affinity and selectivity at M2 and M3 muscarinic receptor subtypes of cyclic and open oxygenated analogs of 4-DAMP. *Farmaco* **1992**, *47*, 1133–47.
(80) Tumiatti, V.; Spampinato, S.; Recanatini, M.; Minarini, A.; Melchiorre, C.; Chiarini, A.; Budriesi, R. Design, synthesis and biological activity of some 4-DAMP-related compounds. 3. *Bioorg. Med. Chem. Lett.* **1995**, *5*, 2325–30.
(81) Ruschig, H.; Korger, G.; Aumuller, W.; Wagner, H.; Weyer, R.; Bander, A.; Scholz, J. New orally effective blood sugar reducing compounds. *Arzneim.-Forsch.* **1958**, *8*, 448–54.
(82) Barrass, B. C.; Elmore, D. T. Formation of cyclic lactams from derivatives of basic amino acids. *J. Chem. Soc.* **1957**, 4830–4.
(83) N-Acylsulfonamides. GB 902881, 1962; Merck & Co., Inc.
(84) Baumann, T.; Baechle, M.; Braese, S. Sulfamidation of 2-Arylaldehydes and Ketones with Chloramine-T. *Org. Lett.* **2006**, *8*, 3797–3800.
(85) Clayton, D. W.; Farrington, J. A.; Kenner, G. W.; Turner, J. M. Peptides. VI. Further studies of the synthesis of peptides through anhydrides of sulfuric acid. *J. Chem. Soc.* **1957**, 1398–407.
(86) Prous. Prous Integrity, Drugs & Biologics, http://integrity.prous.com/.
(87) Tropisetron. Drug Data Rep. 1992, *14*, 863.
(88) Granisetron. Drug Data Rep. 1990, *12*, 519.
(89) Drug Data Rep. 1989, *11*, 530.
(90) Drug Data Rep. 1989, *11*, 896.
(91) Drug Data Rep. 1991, *13*, 469.
(92) Orjales, A.; Alonso-Cires, L.; Lopez-Tudanca, P.; Tapia, I.; Mosquera, R.; Labeaga, L. Benzimidazole-2-carboxylic acid amides and esters: A new structural class of 5-HT3 ligands. *Eur. J. Med. Chem.* **1999**, *34*, 415.
(93) Drug Data Rep. 1992, *14*, 864.
(94) Drug Data Rep. 1992, *14*, 54.
(95) Drug Data Rep. 1988, *10*, 442.
(96) Xue, L.; Godden, J. W.; Bajorath, J. Database Searching for Compounds with Similar Biological Activity Using Short Binary Bit String Representations of Molecules. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 881–886.
(97) Xue, L.; Stahura, F. L.; Godden, J. W.; Bajorath, J. Mini-fingerprints Detect Similar Activity of Receptor Ligands Previously Recognized Only by Three-Dimensional Pharmacophore-Based Methods. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 394–401.
(98) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Model.* **1989**, *29*, 97–101.
(99) Leach, A. R.; Bradshaw, J.; Green, D. V.; Hann, M. M.; Delany, J. J. 3rd. Implementation of a system for reagent selection and library enumeration, profiling, and design. *J. Chem. Inf. Model.* **1999**, *39*, 1161–72.
(100) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic Generation of 3D-Atomic Coordinates for Organic Molecules. *Tetrahedron Comput. Method.* **1992**, *3*, 537–547.
(101) MolNet. Molecular Networks GmbH, http://www.mol-net.de/.
(102) Ihlenfeldt, W. D.; Gasteiger, J. Hash codes for the identification and classification of molecular structure elements. *J. Comput. Chem.* **1994**, *15*, 793–813.
(103) Degen, J.; Rarey, M. FlexNovo: structure-based searching in large fragment spaces. *ChemMedChem* **2006**, *1*, 854–68.

JM0707727