

Chemical Fragment Spaces for de novo Design

Harald Mauser and Martin Stahl*

F. Hoffmann-La Roche Ltd., Pharmaceuticals Division, CH-4070 Basel, Switzerland

Received August 21, 2006

Chemical fragment spaces are combinations of molecular fragments and connection rules. They offer the possibility to encode an enormously large number of chemical structures in a very compact format. Fragment spaces are useful both in similarity-based (2D) and structure-based (3D) de novo design applications. We present disconnection and filtering rules leading to several thousand unique, medium size fragments when applied to databases of druglike molecules. We evaluate alternative strategies to select subsets of these fragments, with the aim of maximizing the coverage of known druglike chemical space with a strongly reduced set of fragments. For these evaluations, we use the Frees fragment space method. We assess a diversity-oriented selection method based on maximum common substructures and a method biased toward high frequency of occurrence of fragments and find that they are complementary to each other.

INTRODUCTION

One of the primary goals of molecular design and cheminformatics is to identify novel chemical entry points for drug discovery programs. The umbrella term “virtual screening” is used to denote in silico compound library searching and filtering procedures in large collections of chemical structures. Even the fastest of these methods, however, can be applied (within a reasonable amount of time) only to a limited number of explicit molecular structures, typically on the order of 10^6 – 10^7 . This is many orders of magnitude lower than the number of potentially interesting druglike organic compounds. Larger collections of molecules can be encoded in the form of chemical fragment spaces.¹ These are sets of fragments of organic molecules combined with connection rules for the generation of explicit molecular structures. A number of de novo design methods based on chemical similarity metrics have been proposed based on chemical fragment spaces. Most of these benefit only from the compact library notation, since they rely on enumerated molecules in an intermediate step.^{2,3} Only methods using an additive molecular similarity metric can fully exploit the potential of large fragment spaces.⁴ This is also the case in structure-based de novo design, where additive scoring functions are used, and fragment spaces have long been employed^{5–8} to build novel molecular structures. The fragments used in this context were typically small, consisting of simple linkers and undecorated ring systems, and connection rules were not used to restrict searches to chemically tractable and druglike compounds.

Here we present strategies to arrive at fragment spaces useful for both similarity-based (2D) and structure-based (3D) de novo design. We start by generating fragments from databases of druglike compounds. Disconnection rules are defined following the RECAP principle,⁹ and the resulting fragments are filtered by means of substructure rules. We then address the question how subsets of these fragment collections can be chosen such that the compounds they represent are diverse and relevant for drug design purposes.

We examine two ways of discarding up to 80% of the fragments and assess the usefulness of these subsets by checking if they encode known druglike compounds or close analogs thereof.

MATERIALS AND METHODS

Compound Collection. We merged compounds from WDI 2004¹⁰ and the Medchem03 databases¹¹ and removed all duplicates. As eventually we were aiming at a template library for de novo design, our focus was on finding novel scaffolds with less emphasis on the details of decoration. Thus we converted all halogens into hydrogen atoms to reduce the number of structurally identical motifs. This resulted in a collection of 85 997 unique compounds represented by SMILES¹² strings.

Fragment Generation. The fragments were generated using a simple iterative disconnection algorithm based on the Daylight toolkit.¹³ The fragmentation rules were coded as Reaction SMARTS¹³ and applied on the canonical SMILES representations of our compound collection. Links marking broken bonds were encoded as “[*m**]” taking advantage of the mass specification *m* in the SMARTS language. In this context, we used *m* as an identifier to differentiate between link types. These links are used as possible attachment points for compatible fragments. Two further procedures were applied to the initial fragment list: First, to allow further functionalization, link atoms were added to those primary and secondary amines that were not derived from fragmentation rules. Second, as ethers are arbitrarily split at one of the C–O bonds, the corresponding alternative alkoxy and alkyl fragments were also generated. The same procedure was applied to amine-derived amine and alkyl fragments. The resulting 24 656 fragments were further subjected to a stepwise filtering procedure to remove large fragments and undesirable motifs. All fragments containing more than three links were removed, as they would lead to overly complex molecules. All fragments containing rings larger than 8 atoms or more than 15 heavy atoms were discarded. Fragments containing metal atoms, obvious reactive centers, or other generally undesirable

* Corresponding author e-mail: martin.stahl@roche.com.














1 Acyl  origin: amides, esters fits 2, 3, 4, 9, 11	2 Amine  origin: amides fits 4, 6, 11, 12, 13	3 Alkoxy  origin: ethers fits 4, 11, 13	4 Acyclic alkyl 1  origin: ethers, amines fits 5, 6, 9, 12, 13, 14
5 Amine  origin: amines fits 4, 13, 14	6 Aminoacyl  origin: ureas fits 2, 4, 9	7 Carbene  origin: alkenes fits 7	8 Acyclic alkyl 2  origin: ring substituents fits 9, 11, 13, 14
9 Aromatic N  origin: heterocycles fits 1, 4, 6, 11	11 (Hetero)aryl  fits 1, 2, 3, 9, 11, 13	12 Sulfonyl  fits 2, 4	13 Heterocyclic C  fits 2, 3, 4, 5, 11
14 Cycloalkyl  fits 4, 5, 8			

Figure 1. Types of fragments used, their origin, and compatibility to other fragments.

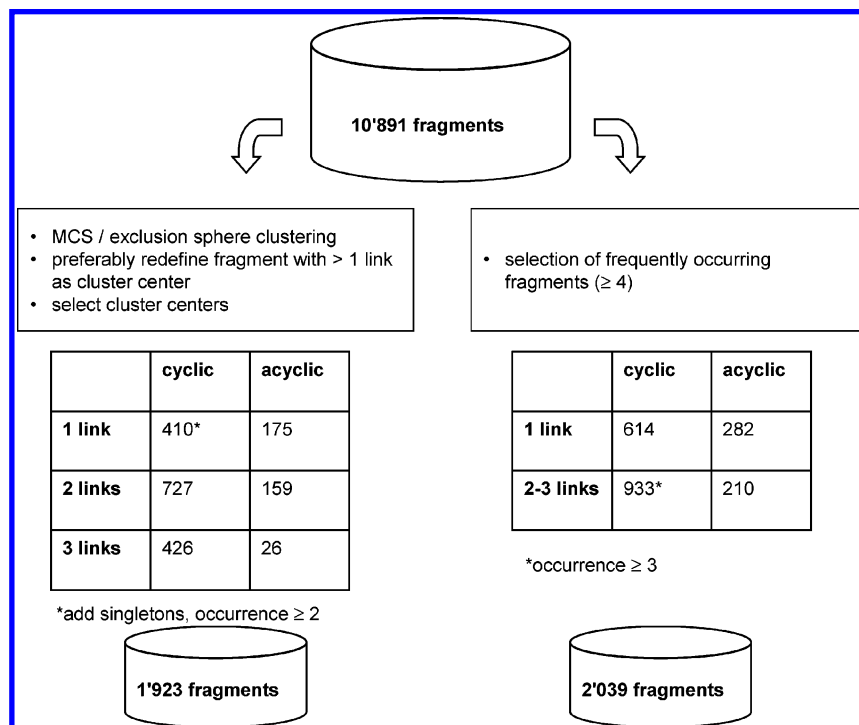


Figure 2. Overview of the two methods to select ~2000 fragments emphasizing diversity (left) and frequency of occurrence (right).

substructures were eliminated by a set of substructure rules. This yielded a final set of 10 891 fragments. Figure 1 gives an overview of the link types and combination rules, which are further discussed in the Results section.

Subset Selection. We compared two techniques of subset generation: selection based on a structural diversity analysis and selection biased toward frequently occurring fragments in the compound collection. For the diversity selection, we employed the maximum common substructure search (MCSS) method. MCSS calculations were done with a previously published algorithm¹⁴ based on the OEChem¹⁵ Python toolkit. We used the scoring scheme as described by Raymond et

al.¹⁶ as a similarity metric. Bond orders and atom types were explicitly considered for molecular graph matching; the atom typing scheme was identical to the one described in ref 17. The MCSS-derived similarity matrix was clustered by means of the exclusion sphere method¹⁸ using an extended algorithm.¹⁴ One modification was introduced: If a cluster center had close neighbors with more links, the neighbor with the larger number of links was redefined as the cluster center, the rationale being that a fragment with a larger number of links is more versatile during de novo design and should thus be preferred. To arrive at a diverse set of approximately 2000 compounds, all cluster centers were selected, but

Table 1. Properties of the Fragment Space and Subsets

	selection method	number of fragments	number of unique rings	log space size		
				≤ 3 frag	≤ 4 frag	≤ 5 frag
1	all fragments	10891	1362 (100%)	11.7	15.9	19.6
2	Random I	2000	438 (32%) ^a	9.5 ^a	12.9 ^a	15.9 ^a
3	Random II	3600	638 (47%)	10.2	14.0	17.2
4	Frequency	2039	273 (20%)	9.5	13.0	15.9
5	Diversity I	1923	530 (39%)	9.9	13.6	16.7
6	Diversity II	3493	874 (64%)	10.3	14.1	17.4
4 ∪ 5	Frequency I + Diversity I	3607	590 (43%)	10.4	14.2	17.6

^a Three random sets of 2000 fragments were chosen. The given values correspond to the mean. Deviations are in the range of 10 unique rings and 0.1 log units.

singletons were omitted. The only exception was the selection for the subset of cyclic building blocks (Figure 2). Here, we chose all cluster centers and added those singleton fragments that were occurring at least twice in the data set. In this fashion, we arrived at 1923 fragments (diversity I). A larger set of 3439 fragments (diversity II) was compiled by adding all the remaining singletons.

Alternatively, we applied a selection procedure biased toward the fragments' frequency of occurrence. All fragments were sorted according to the frequency of their occurrence in the original library and organized into four groups: cyclic and acyclic fragments with one link and cyclic and acyclic fragments with two and three links. To put somewhat more weight on cyclic fragments with more than one link—arguably the most versatile fragments in a drug discovery context—we set the frequency threshold to 3 for this group; otherwise it was set to 4. Figure 2 gives an overview of the set compiled in this manner.

Coverage Calculations. A diverse set of molecules was chosen from the identical collection used for fragment generation. Separately, a diverse set of druglike molecules was chosen from a large set of commercially available compounds. These libraries were filtered (MW: 150–600; number of heavy atoms: 15–40; number of acceptors ≤ 10; PSA ≤ 300; number of rotatable bonds ≤ 7; no rings larger than 8 atoms, substructure filters to remove undesirable motifs). Cluster centers and singletons from exclusion sphere clustering based on Daylight fingerprints (see ref 14 for details) were selected. This resulted in representative sets

of 9470 WDI/Medchem molecules and 10 000 commercially available compounds. Feature Tree fragment space calculations¹⁹ were run for each of these structures as queries and with a target similarity value of 1.0. With this setting, the Ftrees algorithms guarantee the identification of the closest analog encoded in the fragment space according to the Ftrees molecular similarity metric. If the query molecule itself is encoded in the fragment space, it will be found. If not, Ftrees builds close analogs, and the Ftrees similarity metric is likely to generate a number of different structures with comparable similarity values. We chose to generate the top 20 structures with Ftrees and recalculated their similarity to the query molecule with a MOS-based similarity metric¹⁷ taking into account more substructure details. The solution with the highest MOS similarity was chosen as the best match to the query.

RESULTS AND DISCUSSION

The challenge of generating fragment spaces for de novo design is to encode as much information on synthetic feasibility as possible with a set of generic fragments and rules that allow coverage of a large area of chemical space. Two fundamentally different approaches are conceivable: a “forward” approach separately listing synthetically tractable scaffolds and building blocks to be connected in a combinatorial chemistry fashion and a “backward” approach starting from known chemical structures and disconnecting them into fragments to be recombined with a set of chemical rules. We chose this latter approach, as it guarantees a much higher degree of diversity of solutions at the expense of some synthetic tractability. Even here, however, compromises must be found between the extremes of highly restrictive (favoring synthesizability) and highly generic and permissive chemistry rules (favoring diversity). With the Retrosynthetic Combinatorial Analysis Procedure (RECAP) Lewell et al.⁹ introduced a set of disconnection rules that served as a starting point for our work. This is a relatively conservative rule set describing the reverse of simple and widely applicable reactions such as amide formation and coupling of aryl rings. Rings are not opened; small alkyl chains (methyl, ethyl, propyl, isopropyl, butyl) are not cleaved. To arrive at smaller, more versatile fragments, we added two generic rules disconnecting ring systems from acyclic carbon substituents (Figure 1, Materials and Methods). Although the additional

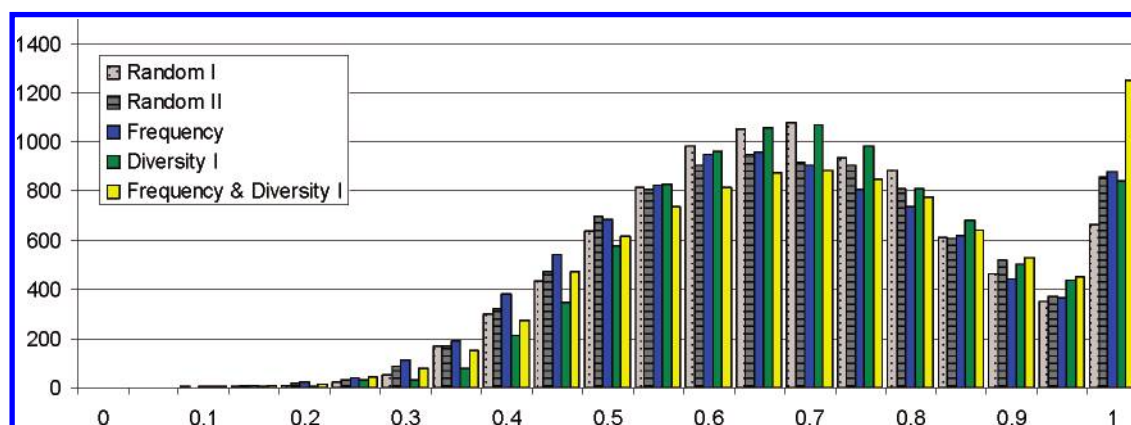


Figure 3. Coverage of chemical space analyzed with query molecules from the WDI-Medchem data set (ca. 10 000 molecules). The histogram shows the distribution of MOS similarity values for the closest analog to each query contained in each of the fragment subspaces (for details of each fragment space see Table 1).

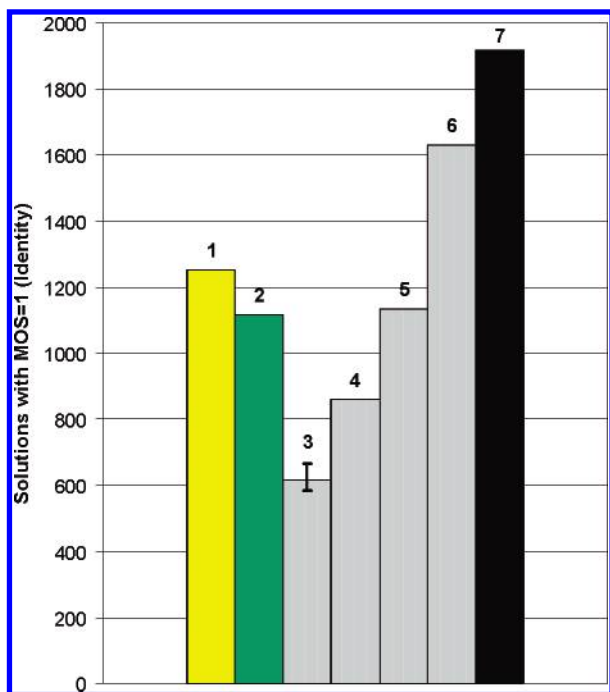


Figure 4. Performance comparison of the combined frequency-biased and diversity fragment sets (gray) to random fragment sets of different size and the full set of all fragments. The plot shows the numbers of solutions structurally identical to the reference molecule. Details about the fragment spaces are listed in Table 1. The bars refer to (1) the combined frequency and diversity I fragment sets (ca. 3600 fragments), (2) diversity II (3) random I (average of three sets with deviation), (4) random II (5) a random set of 5000 fragments, (6) a random set of 8000 fragments, and (7) the full fragment library.

fragment type 14 cannot be obtained by a specific retrosynthetic step, it is well suited to describe the fact that multiple substituents of the same ring system can often be realized through different synthetic routes. This rule set was used to generate fragments from 85 000 compounds from WDI¹⁰ and Medchem libraries.¹¹ Specifically labeled link atoms were used to store the information on the nature of the cleaved bond at each fragment. 85% of the compounds could be cleaved into at least two fragments, and the remaining ones were discarded.

A compatibility matrix of link types determines the “chemistry” encoded in the fragment space (Figure 1). To a

very large extent, these rules determine the size of the fragment space and the nature of its encoded members. Complete compatibility of all fragments with each other would lead to the largest fragment space, while restrictive use—allowing only the reverse of the fragmentation reactions—would ensure the largest degree of synthetic tractability and druglikeness of all its members. We chose an intermediate approach. For example, acyl fragments derived from amides and esters may not only be used to rebuild these functional groups but can also be combined with aryl and alkyl groups to form ketones. On the other hand, only those amine fragments derived from amides—but not those stemming from alkylamines—were allowed to form amides, since experience had shown that this leads to higher chemical tractability of the encoded molecules (at the cost of some duplication of fragments and smaller size of the library).

We then selected subsets of these fragments with three different methods with the aim of evaluating if small sets of fragments would still be able to efficiently cover sizable portions of chemical space. Figure 2 gives an overview on the fragment generation and selection procedure. We chose one substructure-based diverse set, one set containing the most frequently occurring fragments, and one random fragment set. Each of these sets contained approximately 2000 fragments (see Materials and Methods for details).

Table 1 shows some properties of the different fragment spaces. The number of unique ring systems was chosen as an approximate measure of diversity. Of the three sets with approximately equal number of fragments (sets 2, 4, 5), the largest number of unique rings is found in the diversity-based selection. Interestingly, random set 2 contains more unique rings than the set of most abundant fragments 4. Using the Ftrees fragment space program,¹⁹ we calculated the number of molecules encoded by the fragment spaces as a function of the number of fragments per molecule. The diverse fragment subset 5 encodes more than the random set 2 and the frequency set 4. The same is true for the larger diverse set 6 compared to the larger random set 3 and the combination of 4 and 5, but here the difference in size is smaller.

Next, we analyzed the coverage of chemical space by each of these fragment spaces. Approximately 10 000 molecules were selected both from the WDI-Medchem collection and from a collection of commercial vendor catalogs. For each

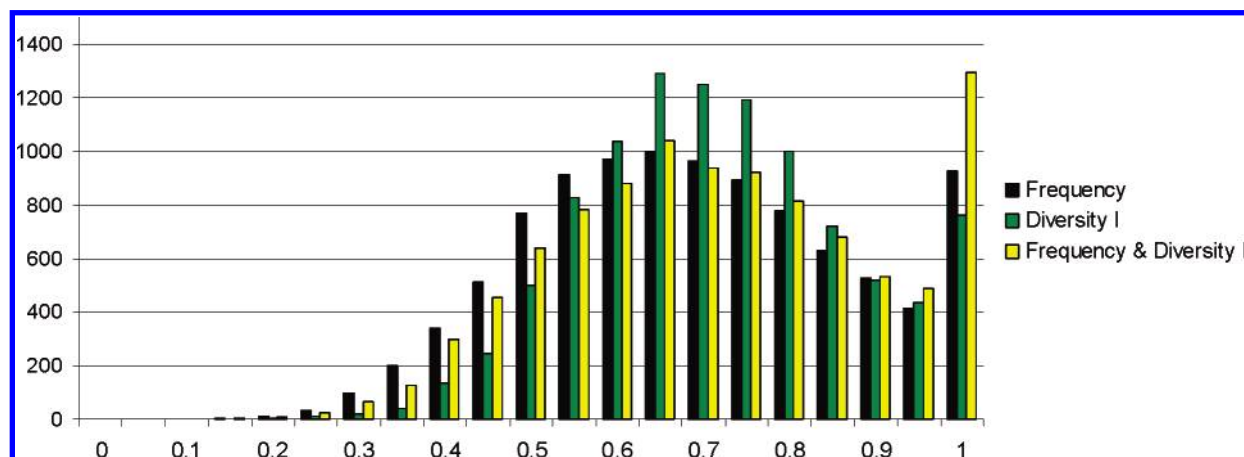


Figure 5. Coverage of chemical space analyzed with query molecules from the commercial vendors data set (10 000 molecules). The histogram shows the distribution of MOS similarity values for the closest analog to each query contained in each of the fragment subspaces (for details of each fragment space see Table 1, the random set is omitted for clarity).

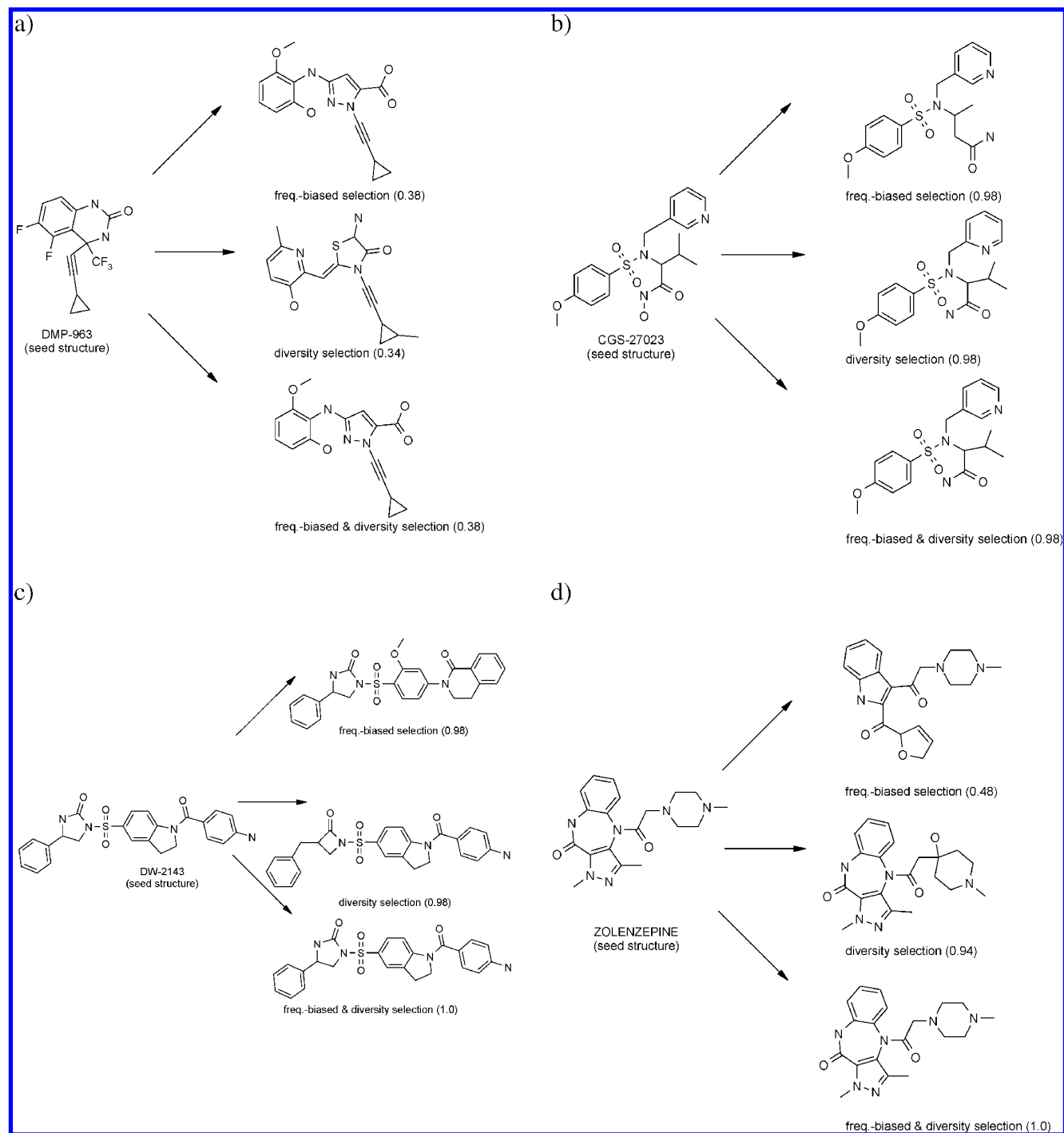


Figure 6. Comparison of top-ranking solutions for the different fragment selections in de novo design. Ftrees was used to generate designs that should match a diverse collection of reference extracted from WDI and Medchem libraries. Examples for seeds where (a) no close analogue was found with any fragment set, (b) all fragments sets arrived at virtually identical solutions, and (c) and (d) only the combined selection could reproduce all features of the seed structure. The given similarity scores in parentheses correspond to the MOS similarities (more information on the seed structures is given in the text).

of these compounds, the Ftrees fragment space method was used to identify the closest analog contained in each of the fragment spaces. A MOS-based similarity metric was employed to assess the similarity between the resulting structures and the query molecules. Low similarity values (<0.4) indicate that no close analog could be generated, whereas high similarities (>0.8) point out that the program could essentially reproduce the query with all its features. Figure 3 shows a distribution of the similarity values for the WDI-Medchem compounds. Defining a MOS similarity

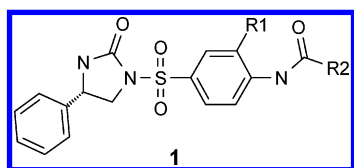
value of 0.8 as a threshold value for a close match, we conclude that every fragment subset covers at least one-fifth of the WDI-Medchem reference compounds. Considering the stringent fragment filtering rules, this is a respectable number. A comparison of the two random selections shows the significant influence of the size of the fragment set on the overall performance. The frequency-biased and the diverse fragment subset perform somewhat better than the random set, but they are not different from each other. Is the method of selecting fragments of minor importance? It rather seems

that 2000 fragments, chosen in whatever fashion, are simply not enough to cover a significant portion of chemical space. A significant improvement in performance is seen when the diverse and frequency-biased fragment sets are combined (and 355 redundant structures are removed). Close to 400 more structures can be reproduced with this set than with a random set of equal size. Figure 4 shows that the combined set even outperforms a random selection containing 1400 more compounds and even the equally sized diverse set (bar 3). This is an unexpected result, as we saw earlier that the size of the selection has a big impact on the number of encoded compounds. It follows that a combination of diversity and frequency of occurrence criteria are well suited to generate compact fragment spaces encoding a diverse set of druglike compounds.

The same analysis was repeated on an independent validation set of commercial compounds, and qualitatively the same results were obtained (Figure 5). The number of query compounds identically contained in the combined fragment set is as high as for the WDI-Medchem data set. This indicates that the fragment space encodes many more compounds relevant for drug design than contained in the parent WDI-Medchem library.

Figure 6 shows a number of examples of top ranking solutions for individual query molecules. In few cases, no appropriate solution was generated (6% for the combined frequency-biased and diversity selection with MOS similarity less than 0.4). This seems to be an acceptable rate corresponding to a small number of difficult design cases consisting of only one very specific fragment. For example DMP-963 (Figure 6a), a HIV-RT inhibitor, is a small molecule with a very particular scaffold that was not present in any of the fragment selections.

The opposite case is CGS-27023, a metalloprotease inhibitor, where we could generate identical chemotypes with all three fragment sets (Figure 6b). In this case, the reference structure is composed of at least three different fragments allowing some variations of the design (the hydroamate fragment was removed during fragment filtering, see Materials and Methods). The de novo results for DW-2143²⁰ (Figure 6c), an oncolytic drug, and Zolenzepine¹⁰ (Figure 6d), a spasmolytic drug, exemplify the complementarity of the frequency-biased and diverse sets. As key fragments with appropriate links were missing in the individual selections, their combination could significantly improve the results. In both cases, all required fragments could only be found in the combined set. Interestingly, the best hits of frequency-based selection for DW-2143 and Zolenzepine are closely related to another oncolytic drug (compound **1**),²¹



or known analgesics,²² respectively. This demonstrates how 2D de novo design can be used to identify novel chemotypes that share the same or at least a similar biological profile. However, analyzing the suitability of fragment spaces for scaffold hopping^{23,24} is beyond the scope of this work and will require further studies.

CONCLUSIONS

We have generated fragment spaces of 2000–4000 members through fragmentation of databases of druglike compounds, filtering, and directed selection of subsets. Our results indicate a combination of the most frequently occurring fragments with a substructure-based diverse subset covers a significantly larger portion of druglike chemical space than the diversity- or frequency-biased subsets alone and are clearly superior to simple random subsets.

We believe that the procedures presented here provide a basis for both 2D and 3D searching and design applications where idea generation for novel chemical entry points is the primary goal and comprehensive listing of closely related alternatives is a burden rather than a desired outcome.

Clearly, both fragment lists and reconnection rules will continue to undergo an evolutionary process. One key area to be addressed in the future is the refinement of the chemical fragmentation and reconnection rules, with the challenging goal of improving both diversity and synthetic tractability of the solutions.

ACKNOWLEDGMENT

We thank all our colleagues in the cheminformatics and molecular modeling group as well as in Discovery Chemistry Basel for stimulating discussions and the support of this work.

REFERENCES AND NOTES

- (1) Lengauer, T.; Lemmen, C.; Rarey, M.; Zimmermann, M. Novel technologies for virtual screening. *Drug Discovery Today* **2004**, *9*, 27–34.
- (2) Naerum, L.; Norskov-Lauritsen, L.; Olesen, P. H. Scaffold hopping and optimization towards libraries of glycogen synthase kinase-3 inhibitors. *Bioorg. Med. Chem.* **2002**, *12*, 1525–1528.
- (3) 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*, 487–494.
- (4) Rarey, M.; Stahl, M. Similarity searching in large combinatorial chemistry spaces. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 497–520.
- (5) Rotstein, S. H.; Murcko, M. A. GroupBuild: A fragment-based method for de novo drug design. *J. Med. Chem.* **1993**, *36*, 1700–1710.
- (6) Boehm, H. J. On the use of LUDI to search the fine chemicals directory for ligands of proteins of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 623–632.
- (7) Boehm, H. J. Site-directed structure generation by fragment-joining. *Perspect. Drug Discovery Des.* **1995**, *3*, 21–33.
- (8) Pearlman, D. A.; Murcko, M. A. CONCERTS: Dynamic connection of fragments as an approach to de novo ligand design. *J. Med. Chem.* **1996**, *39*, 1651–1663.
- (9) 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. Comput. Sci.* **1998**, *38*, 511–522.
- (10) *World Drug Index, Version 2002*; Thomson: Philadelphia, PA, 2002.
- (11) *MedChem03 database*; BioByte: Claremont, CA, and Daylight Chemical Information Systems, Inc.: Aliso Viejo, CA.
- (12) *Daylight Theory Manual*; Daylight Chemical Information Systems, Inc. <http://www.daylight.com/dayhtml/doc/prog/index.html> (accessed Dec 19, 2006).
- (13) *Daylight Toolkit 4.7*; Daylight Chemical Information Systems, Inc.: Aliso Viejo, CA.
- (14) Stahl, M.; Mauser, H. Database Clustering with a Combination of Fingerprint and Maximum Common Substructure Methods. *J. Chem. Inf. Model.* **2005**, *45*, 542–548.
- (15) *OEChem, Version 1.3.3*; Open Eye: Santa Fe, NM.
- (16) Raymond, J. W.; Gardiner, E. J.; Willett, P. Heuristics for similarity searching of chemical graphs using a maximum common edge subgraph algorithm. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 305–316.

- (17) Stahl, M.; Mauser, H.; Tsui, M.; Taylor, N. R. A. Robust Clustering Method for Chemical Structures. *J. Med. Chem.* **2005**, *48*, 4358–4366.
- (18) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747–750.
- (19) Rarey, M.; Dixon, J. S. Feature trees: A new molecular similarity measure based on tree matching. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 471–490.
- (20) Jung, S.-H.; Lee, H.-S.; Song, J.-S.; Kim, H.-M.; Han, S.-B. et al. Synthesis and antitumor activity of 4-phenyl-1-arylsulfonyl imidazolidinones. *Bioorg. Med. Chem. Lett.* **1998**, *8*, 1547–1550.
- (21) Kim, S.; Park, J. H.; Koo, S.-Y.; Kim, J. I.; Kim, M.-H. et al. Novel diarylsulfonylurea derivatives as potent antimitotic agents. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 6075–6078.
- (22) Bell, M. R. 3-Carbonyl-1-aminoalkyl-1H-indoles useful as analgesics. *Eur. Pat. Appl.* 1986; EP 171037.
- (23) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-Hopping" by topological pharmacophore search: a contribution to virtual screening. *Angew. Chem. Int. Ed.* **1999**, *38*, 2894–2896.
- (24) Boehm, H.-J.; Flohr, A.; Stahl, M. Scaffold Hopping. *Drug Discovery Today Technol.* **2004**, *1*, 217–224.

CI6003652