Molecular Shape Diversity of Combinatorial Libraries: A Prerequisite for Broad Bioactivity†

Wolfgang H. B. Sauer and Matthias K. Schwarz*

Serono Pharmaceutical Research Institute, Department of Chemistry, 14, chemin des Aulx, 1228 Plan-les-Ouates, Geneva, Switzerland

Received September 4, 2002

A computational method to rapidly assess and visualize the diversity in molecular shape associated with a given compound set has been developed. Normalized ratios of principal moments of inertia are plotted into two-dimensional triangular graphs and then used to compare the shape space covered by different compound sets, such as combinatorial libraries of varying size and composition. We have further developed a computational method to analyze interset similarity in terms of shape space coverage, which allows the shape redundancy between the different subsets of a given compound collection to be analyzed in a quantitative way. The shape space coverage has been found to originate mainly from the nature and the 3D-geometry (but not the size) of the central scaffold, while the number and nature of the peripheral substituents and conformational aspects were shown to be of minor importance. Substantial shape space coverage has been correlated with broad biological activity by applying the same shape analysis to collections of known bioactive compounds, such as MDDR and the GOLD-set. The aggregate of our results corroborates the intuitive notion that molecular shape is intimately linked to biological activity and that a high degree of shape (hence scaffold) diversity in screening collections will increase the odds of addressing a broad range of biological targets.

INTRODUCTION

Combinatorial Chemistry enables the expedient, simultaneous synthesis of large compound collections ("*libraries*"), by bringing together sets of reactive monomers ("*building blocks*") $A_1,A_2...A_m$; $B_1,B_2,...B_n$; ··· $Z_1,Z_2...Z_p$, in such a way that in principle all combinations of final compounds of general structure AB···Z. are formed.[1−3] In most cases, one building block remains constant throughout all compounds of a given library; it is then referred to as "*template*", "*scaffold*", or "*chemotype*" (for a more detailed definition of the template concept, see ref 4). Toward the end of the past decade, there has been a paradigm shift in the way that the pharmaceutical industry has applied combinatorial chemistry to drug discovery. Library synthesis for primary screening has steered away from numerically large libraries (>10 000 members) of a limited number of scaffolds and instead moved toward collections of small libraries (<1000 members) comprising many different chemotypes. From a practical point of view, this trend is quite remarkable, as experience shows that the most time-consuming step in combinatorial library production typically consists of validating and optimizing the chemistry for each given scaffold-type. Therefore, the generation of several small libraries around different scaffolds will usually require much more time and resources than the assembly of one large library based on a single chemotype. Clearly, practicality and feasibility arguments must have been superseded by other considerations. A common (yet rarely publicized) observation in high-throughput screening of early combinatorial libraries against a number of biological targets was that, for a given scaffold, the biological results tended to be "sporadic", i.e. the hit rate throughout the different targets was either very high or essentially zero. On the other hand, compound sets containing a wealth of different chemotypes, such as natural product or commercial compound collections, tended to exhibit more consistent hit rates across a variety of targets. These observations led us and others[5] to speculate that a combinatorial library derived from a single scaffold, irrespective of the library size, could innately not become diverse enough to be able to interact with a number of different biological targets. This could intuitively be explained by the preset 3D geometry of a given scaffold restraining the number of possible spatial arrangements of the peripheral substituents, hence limiting the number of attainable molecular shapes to fit the relevant cavities presented by the biological targets. In this report, we attempt to formally verify this hypothesis and show that small multiple-scaffold libraries are superior to large single-scaffold libraries in terms of their biorelevance, i.e. their potential to hit a broad panel of biological targets. To this end, we propose a novel computational approach allowing compound collections to be rapidly assessed and compared with respect to their "shape diversity".

METHODOLOGY

Molecular Descriptors: Normalized PMI Ratios (NPRs). Any attempt at assessing and comparing the diversity of compound sets must first face the challenge of defining which molecular properties or descriptors the diversity calculations

* Corresponding author phone: +41 22 706 9820; fax: +41 22 794 6965; e-mail: Matthias.Schwarz@Serono.com.

† Dedicated to the memory of Dr. Nabil El Tayar.

10.1021/ci025599w CCC: $25.00
Published on Web 03/14/2003

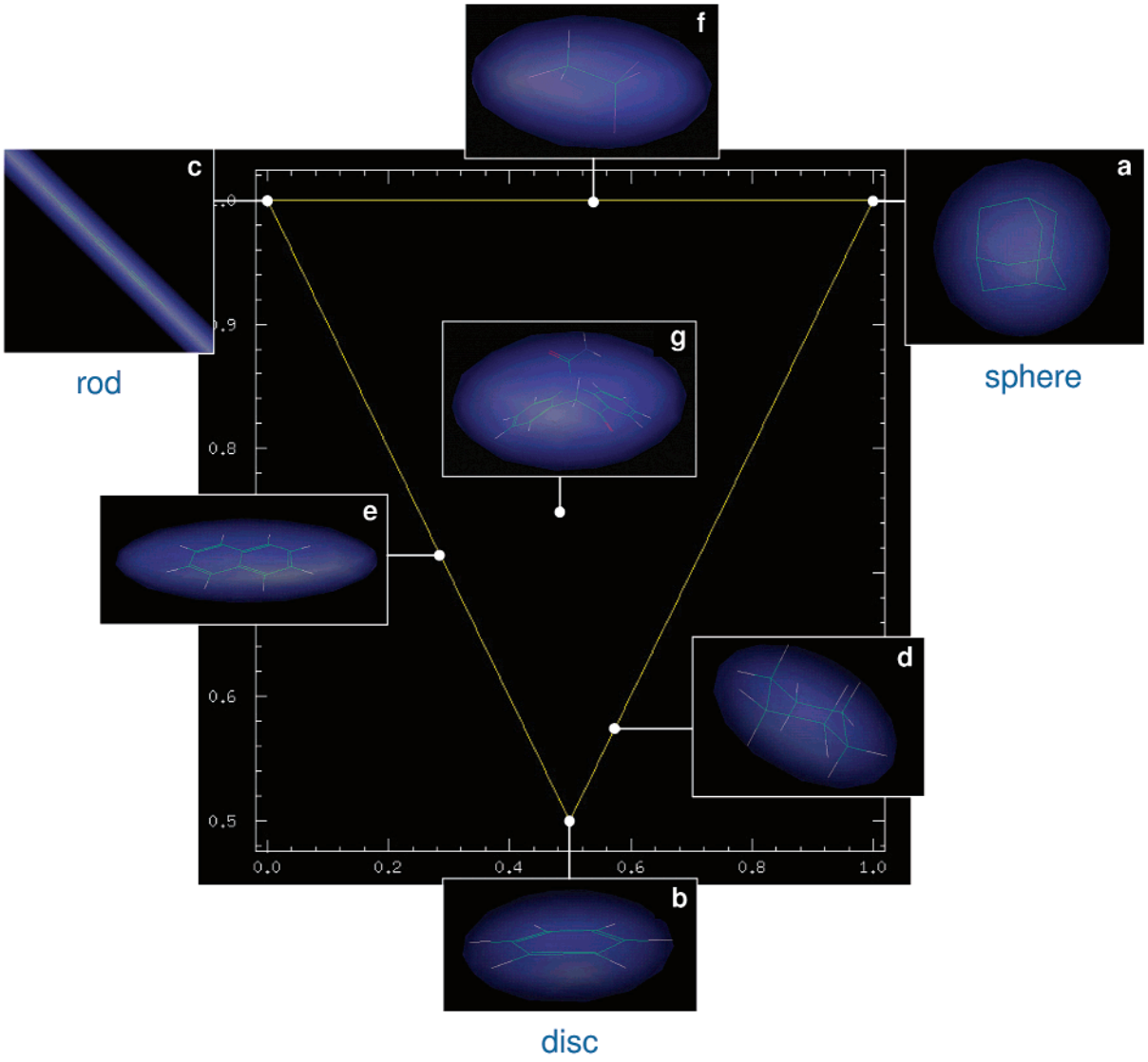
should be based upon. Diversity (or dissimilarity) and its antipode similarity are meaningful attributes only with respect to specified molecular descriptors, since every given set of descriptors generates a unique chemistry space with an idiosyncratic distribution of the compounds to be compared.[6] From the wealth of literature published on the subject[7−28] one can safely conclude that, to date, there is no such thing as a generally applicable "default set" of molecular descriptors,[29] and the choice of the "right" descriptors will usually be governed by the nature of the questions to be answered.

The first goal of the present study was to provide formal evidence that small multiple-scaffold libraries are indeed superior to large single-scaffold libraries in terms of biorelevant diversity, i.e. the potential to produce hits against a panel of biological targets. We reasoned that in order to provide a—chemically intelligible—solution to this problem, a molecular descriptor or property needed to be (i) a priori correlated with, and predictive for, biological activity, (ii) (back-)translatable into chemical structure terms, and (iii) fast to calculate. Most descriptors shown to correlate with biological activity are derived a posteriori by training an initial descriptor set using a series of compounds with associated biological data on a given target or target class (for a nice example, see refs 25 and 26). With respect to the evaluation of the biorelevant diversity of primary screening libraries, these trained descriptors are less suitable, because typically, at the library generation stage, no specific biological target will be known. Some descriptor sets have been described, which appear to be a priori correlated with, thus predictive for, biological activity against a range of targets (see ref 7). However these tend to be less chemically intuitive, meaning that it will be difficult for the chemist to translate and use the information as guidance for library synthesis. Our attention was drawn to an original approach by the Vertex group, in which bioactive compounds, in the specific case known drugs from the CMC-database, were analyzed with respect to their two-dimensional molecular frameworks,[30] resulting in the concept of classifying drugs by common "shape themes".[31] Although the term "shape" in these reports was used as a synonym for two-dimensional topological graphs, it instigated us to consider three-dimensional shape as our molecular descriptor of choice. Three-dimensional molecular shape intuitively meets the above criterion of a descriptor being a priori correlated with, and predictive for, biological activity, simply because a compound will only modulate the activity of a biological target, if its 3D-shape can match the appropriate cavities, clefts, or bulges presented by the biological counterpart. To the chemist, molecular shape is an intelligible descriptor that can be rationalized and, if needed, modified in a predictable way. As to the third criterion, the time needed to compute the molecular shape of a given compound will depend on the way in which the shape information is captured and represented. A number of shape descriptors of varying complexity have been published, including Meyer's Globularity,[32] Kier's κ-indices,[33] Jurs' Charged Partial Surface Areas,[34] and Shadow Projections,[35] catShape descriptors,[36] 3D autocorrelograms,[37] and more.[38] Other methods exist to provide shape similarity measures for pairs of compounds,[39] but for the comparison of whole libraries these appear much less suited. As a compromise between complexity and information content of the representation, we decided to evaluate normalized ratios of principal moments of inertia (PMI) for their potential to serve as an intuitive, albeit rather crude, way to describe molecular shape. PMIs on their own, derived either computationally or experimentally from IR or microwave spectra, have previously been used to assess molecular properties such as shape, geometry, and conformational parameters.[40] PMIs are conveniently accessible from many software packages, e.g. Cerius^2,[41] MOE,[42] Sybyl,[43] Tsar,[44] and others. Thus, compound libraries were enumerated in 2D using Cerius^2, and for each compound a 3D structure was derived, using, in a first instance, Corina,[45,46] which was then used to calculate the three principal moments of inertia, sorted by ascending magnitude $\mathbf{I_1}$, $\mathbf{I_2}$, and $\mathbf{I_3}$. Subsequently, normalization was performed by dividing the two lower PMI-values ($\mathbf{I_1}$ and $\mathbf{I_2}$) by the highest value ($\mathbf{I_3}$), generating two characteristic values of normalized PMI ratios (NPRs) for each compound ($\mathbf{I_1/I_3}$ and $\mathbf{I_2/I_3}$). This completely eliminates the dependency of the chosen representation on the size of the molecules under investigation, which eases the need for decorrelation procedures when used in combination with other descriptors such as molecular weight, volume, or surface area. Furthermore, due to the intrinsic characteristics of the inertia tensor, the following relation will be fulfilled:

$$\mathbf{I_2/I_3} \geq \max(\mathbf{I_1/I_3}, 1\text{-}\mathbf{I_1/I_3}) \quad (1)$$

Therefore, when finally plotted against each other, the resulting graph, shown in Figure 1, represents an isosceles triangle, into which all compounds are projected. It is defined by its three corners, wherein the vector [$\mathbf{I_1/I_3}$, $\mathbf{I_2/I_3}$] equals [1,1], [0.5,0.5], and [0,1], corresponding to archetype "envelope" shapes of, respectively, spheres, disks, and rods. Some examples for molecules exhibiting these extremes as well as intermediate geometries are also shown in Figure 1.

Combinatorial Libraries (A), (B), (C), and S_1-S_{50}: Three Points of Diversity. With this computational approach in hand, we turned our attention to devising an appropriate chemical model case related to the initial question about the relative merits of multiple-scaffold versus single-scaffold libraries. Looking first at a series of single compounds built up from different scaffolds but the same set of three peripheral building blocks R1−R3, it is obvious that the resulting molecular shapes differ to a great extent (see Figure 2). Intuitively, this is clear, because each scaffold has defined exit vectors for its substituents and hence will orient them differently in three-dimensional space. The main question, however, is whether the apparent lack of diversity in terms of accessible shapes due to the fixed geometry of one given scaffold can be compensated by simply increasing the number of peripheral building blocks, that is by making a single-scaffold library sufficiently big. To investigate this, we took the example of a benzodiazepine scaffold with three positions for peripheral substituents. For the sake of simplicity, all three substituents were chosen from the same set of 13 diverse R-groups (see Table 1 and Figure 3), leading to a relatively small library of 2197 members (**A**). This library was then enlarged to >100 000 members in two ways: (i) by increasing the set of peripheral substituents from 13 to 50 (while maintaining the same benzodiazepine scaffold) and (ii) by increasing the number of central scaffolds from 1 to 50 (while keeping the initial set of 13 peripheral substituents),



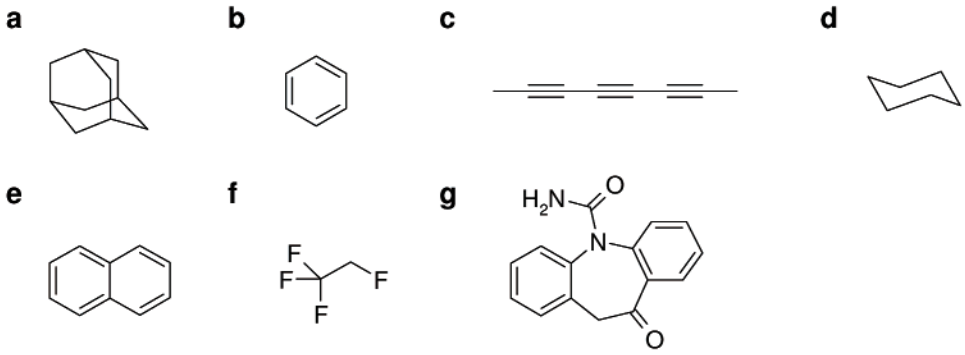


Figure 1. Normalized PMI ratios as shape descriptors: the position within the triangle reveals the "envelope shape".

resulting in two big libraries **(B)** and (**C**), of 125 000 and 109 850 members, respectively (see Table 1). The set of 50 peripheral substituents as well as the 13-member subset thereof were chosen with a view to maximum diversity in terms of functional groups and steric factors (see Figure 3). The scaffolds were chosen from reports on published combinatorial library syntheses (see Figure 4).[47−50] Finally, all three libraries were subjected to the shape diversity analysis outlined in the first part of this section.

Combinatorial Libraries S_{51}−S_{65}: Two Points of Diversity. Two-point diversity libraries, denoted as S_{51}−S_{65}, were constructed from the minimal scaffolds A_{51}−A_{65} (see Figure 5) and the 50 peripheral substituents B_1−B_{50} (see Figure 3), according to Table 1, resulting in 15 2500-member libraries. For the sake of consistency, the same set of diverse peripheral substituents was used as in the case of the three-point diversity libraries, chosen with a view to maximum diversity in terms of functional groups and steric factors.

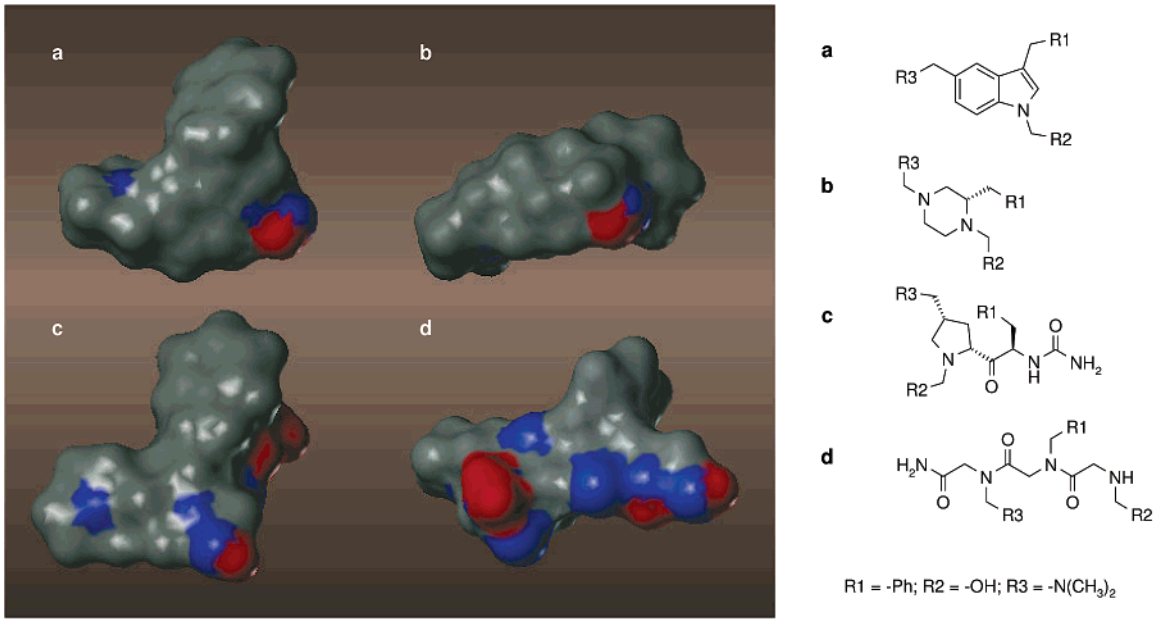


Figure 2. Shape comparison of single compounds originating from different scaffolds and identical peripheral building blocks (R1-R3).

Table 1. Scaffold and Peripheral Substituent Combinations Used for Library Construction[a]

library	scaffolds (see Figures 4 and 5)	peripheral substituents R (see Figure 3)
(**A**) = **S**$_1$	A_1	B_5, B_6, B_{11}, B_{18}, B_{27}, B_{29}, B_{32}, B_{41}, B_{43}, B_{45}, B_{47}, B_{48}, B_{50}
(**B**)	A_1	B_1-B_{50}
(**C**)	A_1-A_{50}	B_5, B_6, B_{11}, B_{18}, B_{27}, B_{29}, B_{32}, B_{41}, B_{43}, B_{45}, B_{47}, B_{48}, B_{50}
S$_n$ [$n = 2-50$]	A_n	B_5, B_6, B_{11}, B_{18}, B_{27}, B_{29}, B_{32}, B_{41}, B_{43}, B_{45}, B_{47}, B_{48}, B_{50}
S$_m$ [$m = 51-65$]	A_m	B_1-B_{50}

[a] See Figures 3−5.

All libraries were subjected to the shape diversity analysis outlined in the first part of this section.

Similarity Indices. To assess the similarity between libraries in a quantitative manner, cell-based indices have been defined in analogy to the Carbo indices[51] and Hodgkin indices[52] traditionally used to quantify the similarity between pairs of molecules.[53] First, cells have been constructed by subdividing the triangle area into 2500 nonoverlapping, isosceles triangles of equal size.

Conventionally, the Carbo index R and the Hodgkin index H between two molecules A and B are computed from the respective formulas

$$R_{AB} = \frac{\sum_{i=1}^{N}(P_A P_B)}{\sqrt{\sum_{i=1}^{N} P_A{}^2}\sqrt{\sum_{i=1}^{N} P_B{}^2}}$$

and

$$H_{AB} = \frac{2\sum_{i=1}^{N}(P_A P_B)}{\sum_{i=1}^{N} P_A{}^2 + \sum_{i=1}^{N} P_B{}^2}$$

where N = the total number of grid points and P = the property being compared at the ith grid point. In the present case, N = the total number of cells, A and B, respectively, denote the two libraries to be compared, and P is one of the following properties: the count of compounds of a library in the ith cell, the fraction of compounds of a library in the ith cell,[54] or the difference between the actual number of compounds in the ith cell and the number of compounds a cell would contain, if the library was distributed homogeneously across all cells.

In all cases related to this work, all indices were found to be highly correlated among each other ($r = 0.84-1.00$, $p < 0.001$), consequently only Carbo indices will be reported in the following sections.

RESULTS AND DISCUSSION

The shape triangle diagrams obtained from the analysis of the three libraries (**A**), (**B**), and (**C**) described in the previous section are shown in Figure 6. For consultation of the individual shape triangle diagrams of the 50 sublibraries of library (**C**), **S**$_1$−**S**$_{50}$, based on the 50 scaffolds **A**$_1$−**A**$_{50}$ (see Figure 4), the reader is referred to Chart 1, Supporting Information.

Inspection of the triangle diagram of library (**A**) reveals that the 2197 benzodiazepine compounds are spread in an intermediate region pointing to composite molecular envelope shapes with some spherical, some elongated, and some discoid character. Notably though, some areas of the triangle remain entirely unpopulated, such as the region around the right-hand corner and a strip along the left-hand edge,

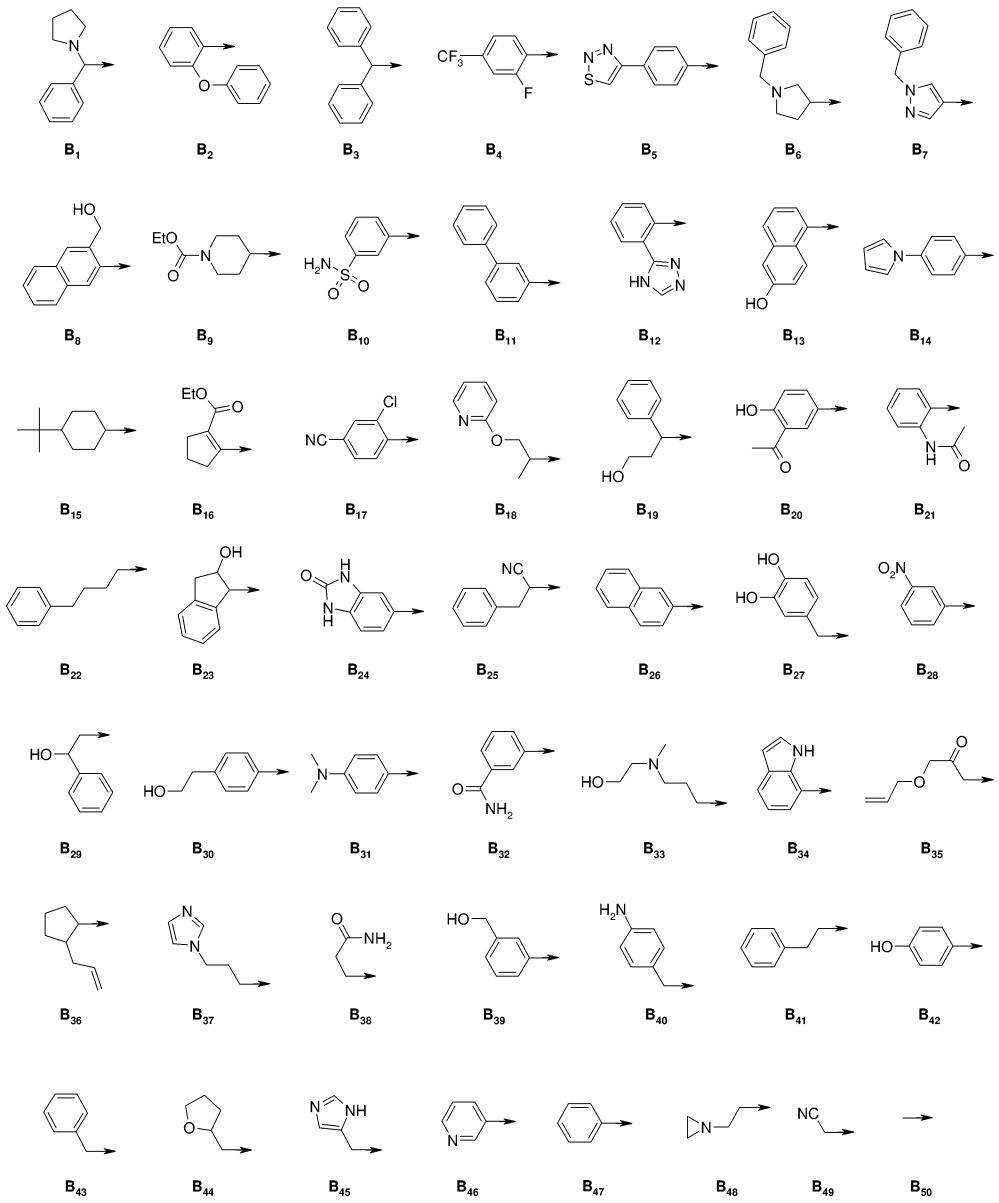


Figure 3. Peripheral substituents, $\mathbf{B_1}-\mathbf{B_{50}}$, used for library construction according to Table 1. Arrows denote points of attachment to the scaffolds, $\mathbf{A_1}-\mathbf{A_{65}}$ (see Figures 4 and 5).

indicating the absence of, respectively, highly spherical and elongated discoid shapes. When the library size is multiplied by a factor of >50 by augmenting the number of peripheral substituents, such as in library (**B**), the overall spread partly improves, with the area corresponding to shapes with predominantly spherical character now becoming populated as well. Intriguingly, however, the left-hand strip remains completely unoccupied, suggesting that, based on the specific benzodiazepine scaffold in question, elongated flat molecular shapes can *intrinsically* not be accessed, regardless of the library size. Indeed, this was shown to hold true even for library sizes exceeding by far 1 000 000 members (data not shown, but see discussion section on multiconformation analyses). If, on the other hand, the single-scaffold library (**A**) is extended to a multiscaffold collection of small libraries, such as library (**C**), a nearly complete coverage of the shape triangle diagram is observed, despite a slightly lower total number of compounds compared to library (**B**). In particular the stretch along the left-hand edge, found to be "inaccessible" to the benzodiazepine library (**B**), is densely populated in the case of library (**C**), indicating a high incidence of elongated flat shapes. While these results indicate that multiscaffold libraries are indeed more diverse in terms of molecular shape than single-scaffold libraries of comparable size, they also raise a number of follow-up questions. For example, considering the remarkable pharmaceutical track record of benzodiazepines, one can rightfully ask whether the empty area in the triangle diagram of library (**B**) has any biological relevance at all. It can further be questioned if and to which extent the results have been distorted by restricting the shape analysis to only one conformation per compound (i.e. the Corina structure). Finally, the dense distribution observed in the diagram of library (**C**) points to a possible redundancy between some of the 50 sublibraries in terms of shape diversity, implying that the number of scaffolds may be significantly reduced, without impairing the overall coverage of the shape diagram.

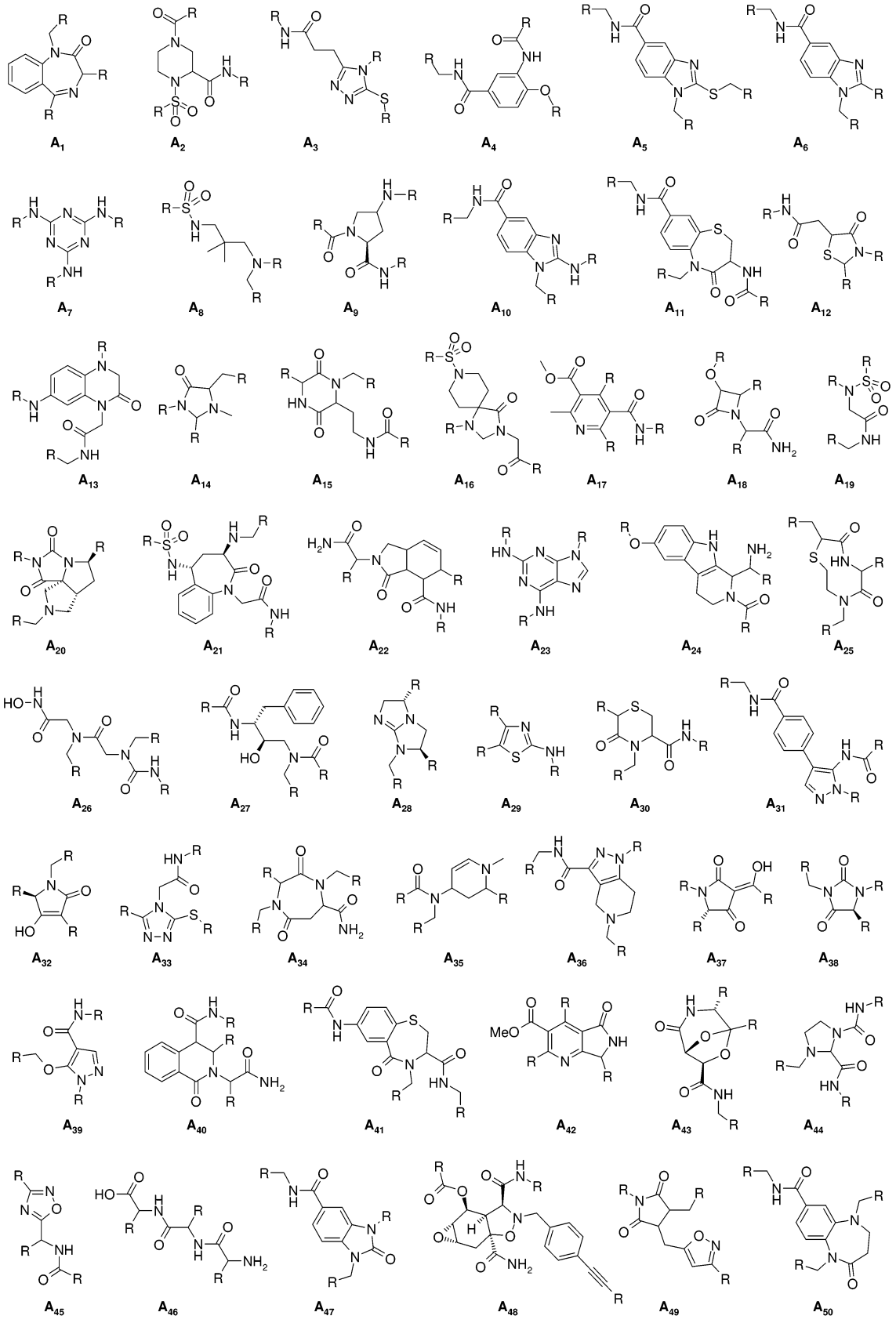


Figure 4. Three-point diversity scaffolds, $\mathbf{A_1}-\mathbf{A_{50}}$, used for library construction according to Table 1. The symbol R specifies the positions of peripheral substituents, $\mathbf{B_1}-\mathbf{B_{50}}$ (see Figure 3).

The multiscaffold library (**C**) differs from the single-scaffold library (**B**) in terms of molecular shape, in that it contains more compounds with markedly discoid and elongated shape (see Figure 6). But is this additional shape range pharmacologically meaningful, i.e. prone to translate also into additional biological activities? To investigate this,

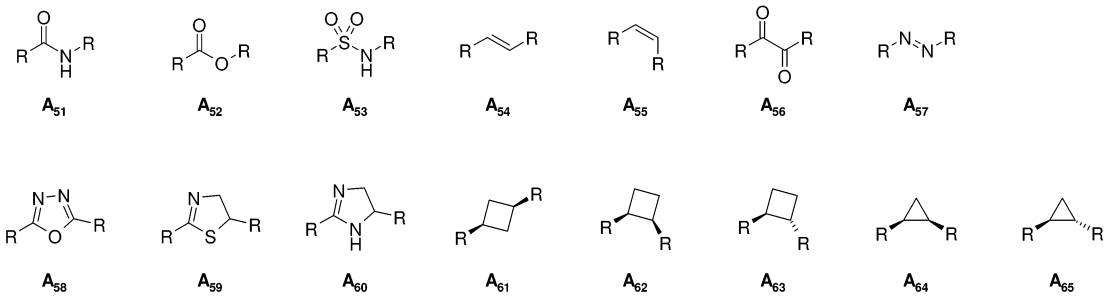


Figure 5. Two-point diversity minimal scaffolds, $\mathbf{A_{51}}-\mathbf{A_{65}}$, used for library construction according to Table 1. The symbol R specifies the positions of peripheral substituents, $\mathbf{B_1}-\mathbf{B_{50}}$ (see Figure 3).

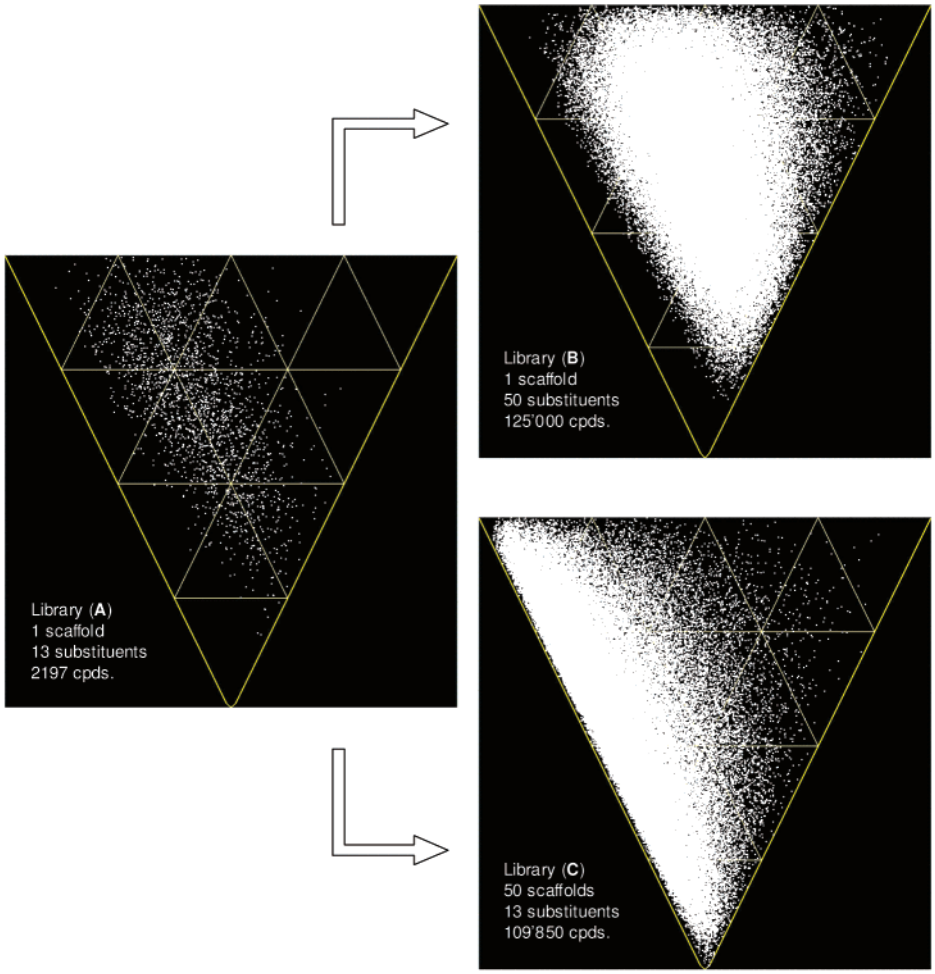


Figure 6. Comparison of libraries (**A**), (**B**), and (**C**) based on shape diversity.

we decided to apply our shape diversity analysis approach to two collections of compounds known to interact with biological targets, namely the MDL Drug Data Report (MDDR), comprising 101 800 compounds currently in development as potential drugs, and the GOLD set, containing 136 three-dimensional compound structures as derived from crystal structures of small-molecule/protein complexes.[55] To this end, the 2D-structures of the 98 550 chemical structures in the MDDR were first transformed into the corresponding Corina 3D-structures, in the same way as before the compounds of libraries (**A**), (**B**), and (**C**). As to the GOLD-set, the 3D-structures were directly obtained from the given source[55] and used without further modification other than the addition of hydrogen atoms. The resulting triangle diagrams, shown in Figure 7, reveal that for both collections, a considerable proportion of the compounds is scattered within the critical area along the left-hand triangle side, demonstrating that flat elongated molecular shapes are by no means incompatible with biological activity, rather on the contrary. The remarkable similarity of the plots derived from library (**C**) and MDDR gives an insight into why compound collections containing multiple chemotypes have frequently been found to outperform much larger (but single-scaffold) combinatorial libraries in terms of their average hit-rates against different biological targets. On the other hand, the initially mentioned sporadic ("all-or-nothing") hit-rates often observed with early single-scaffold combinatorial libraries can equally be rationalized, based on the shape diversity diagrams obtained for libraries (**A**) and (**B**): for biological targets with cavities matching only flat elongated

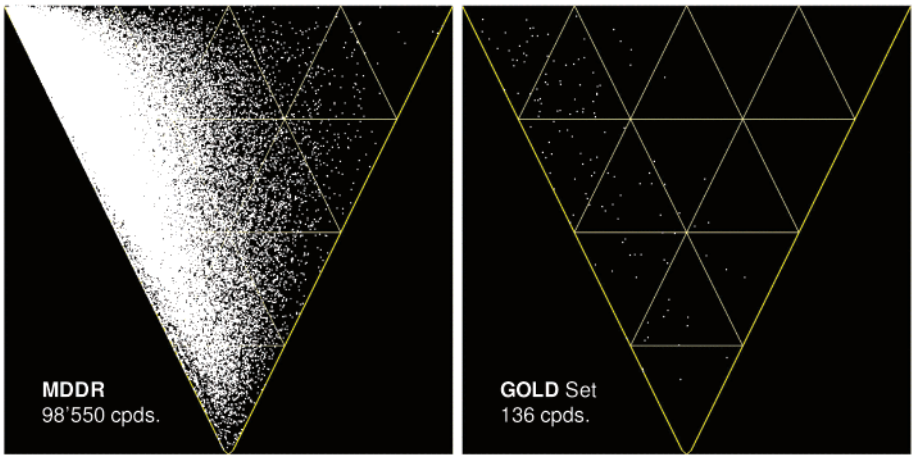


Figure 7. Pharmacological relevance of maximum shape triangle coverage: Analysis of collections containing known bioactive compounds (MDDR and GOLD-set).

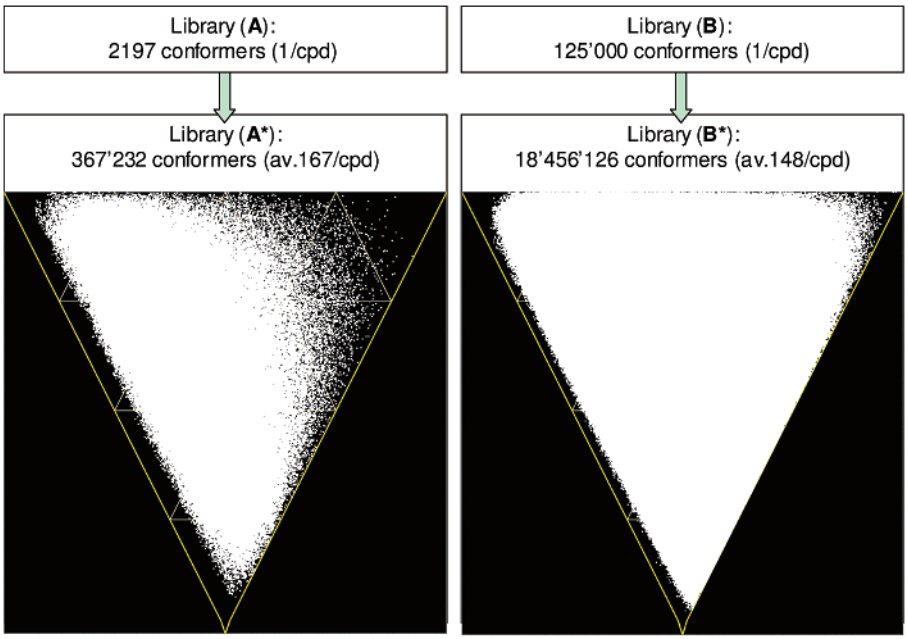


Figure 8. Multiconformation shape analysis: libraries (**A***) and (**B***).

molecular shapes, the chances of finding a hit in those libraries are essentially zero, irrespective of the library size. Conversely, for targets requiring more globular shapes for optimal interaction, corresponding to the densely populated area of the triangle diagram, the likelihood of identifying hits from a collection like library (**B**) is significant.

The shape distribution within the GOLD set (see Figure 7) is qualitatively very similar to that observed for MDDR and library (**C**), revealing again a considerable number of flat and elongated molecular shapes. This observation is of particular importance, because the 3D-structures of the GOLD-compounds undeniably reflect the reality as encountered in an authentic biological context, meaning that the end results do not depend on the choice and quality of an initial 2D/3D-structure conversion step. Taking this argument one step further, however, one may now question whether the results of the other compound collections, in particular the "forbidden" areas observed in the shape diagrams of libraries (**A**) and (**B**) (see Figure 6), are not a mere artifact caused by limiting the shape analysis to the Corina structures, i.e., to only one conformation per compound. Corina 3D-structures are based on standard bond-lengths, bond-angles, and ring-conformations, taking into account atom type, hybridization state, and bond order. In a validation study comparing several hundred small-molecule X-ray structures from the CSD (Cambridge Structural Database)[56] with the corresponding 3D structures generated by Corina, the program was found to predict with high accuracy.[57,58] Considering, however, that the 3D-structure determined from crystals of the small-molecule alone is not necessarily congruent with that adopted in a complex with the biological target,[59] we nevertheless decided to repeat the analysis of libraries (**A**) and (**B**), this time, however, taking molecular flexibility into account. For this purpose up to 250 diverse conformations were generated for each molecule using the program catConf,[60] allowing even for highly unlikely conformers of up to 20 kcal higher energy. For each conformer of each compound, NPRs were calculated as above. The results, presented in Figure 8, show that even when including an average of 148 conformations per compound, the picture globally remains the same as described before (see Figure 6 and text), in that the critical region along

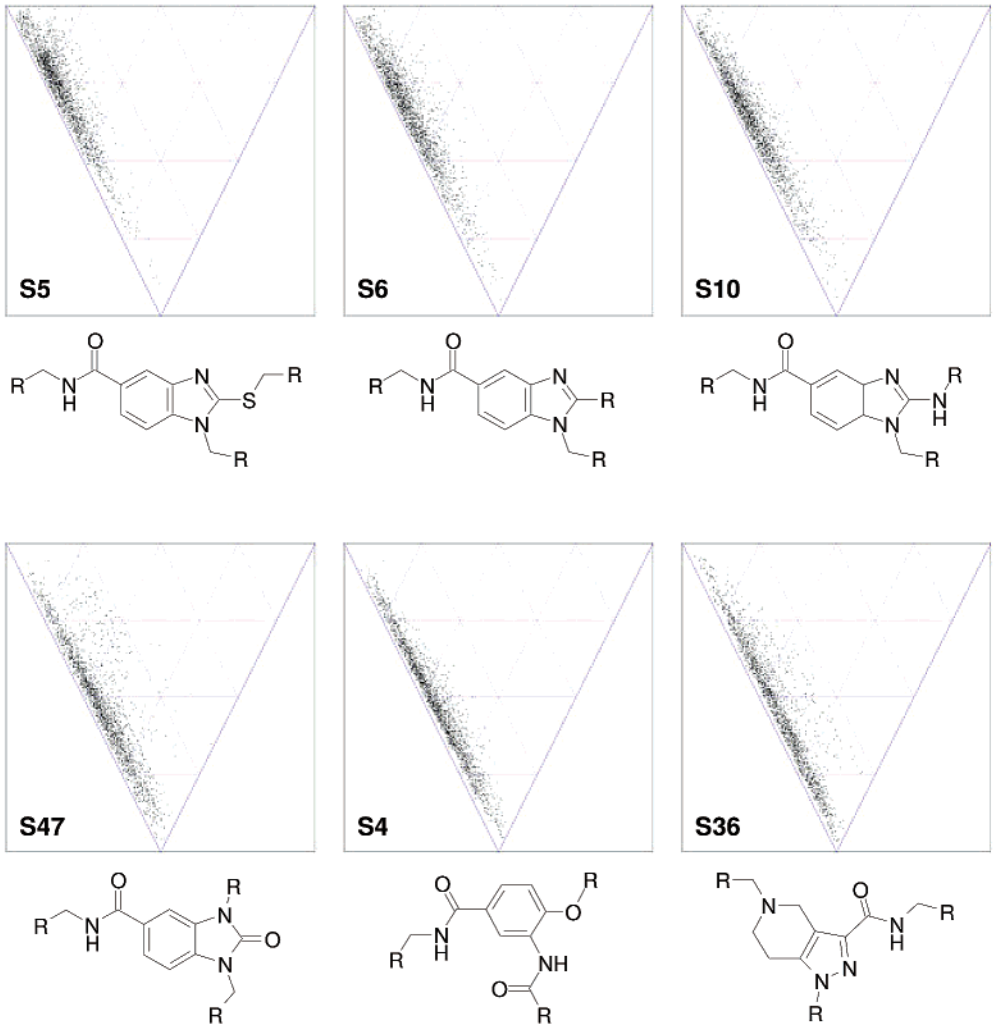


Figure 9. Similarity and dissimilarity among sublibraries with respect to envelope shape and 2D-structure.

the left-hand triangle edge is again found to remain empty for both single-scaffold libraries.

Taken all together, the data presented so far strongly suggest that in order to generate a compound library of maximum biorelevant diversity, the best strategy is to produce numerous small sublibraries based on many different chemotypes. However, as mentioned in the introductory remarks, this puts considerable strain on the practicing chemist, because the time and effort spent on library generation tend to increase proportionally with the number of different scaffold chemistries, in contrast to just enlarging a preexisting library by adding some more peripheral substituents, which is usually straightforward. It is therefore crucial to prioritize the different sublibraries against each other, based on their relative contributions to the overall shape diversity of the collection. Based on these considerations and on the apparent shape redundancy within library (**C**) (see Figure 6), we set out to explore whether the shape diversity analysis could potentially serve as a tool to decide which libraries to synthesize first and which ones at a later stage. Mere visual inspection of the shape diagrams associated with the different sublibraries of library (**C**) (see Chart 1, Supporting Information) reveals that there are indeed quite a few with very similar distribution patterns. From the diagrams shown in Figure 9, one may, for example, recommend to select only one out of the three benzimidazole libraries ($\mathbf{S_5}$, $\mathbf{S_6}$, $\mathbf{S_{10}}$) for a first cycle of library production. This example may seem trivial, since the recommendation derived from the shape diversity analysis (gratifyingly) coincides with what one would intuitively propose by simply looking at the 2D-structures of the three scaffolds $\mathbf{A_5}$, $\mathbf{A_6}$, and $\mathbf{A_{10}}$. In other cases, the computationally derived similarity in shape is much less obvious based solely on the 2D-structures. Thus, judging from the shape distribution patterns, sublibraries $\mathbf{S_{47}}$, $\mathbf{S_4}$, and $\mathbf{S_{36}}$ appear similar to each other but different from sublibraries $\mathbf{S_5}$, $\mathbf{S_6}$, and $\mathbf{S_{10}}$, in that their molecular envelope shapes have a higher discoid and less elongated character. In 2D-structural terms, the difference between the two groups can best be understood by focusing on the benzimidazolone scaffold $\mathbf{A_{47}}$, which, in contrast to the benzimidazole scaffolds $\mathbf{A_5}$, $\mathbf{A_6}$, and $\mathbf{A_{10}}$, orients one substituent *away* from the principal molecular rotation axis, thereby increasing the molecular extension into the second dimension. The shape similarity within the group $\mathbf{S_{47}}$, $\mathbf{S_4}$, and $\mathbf{S_{36}}$, which is a priori less apparent from the 2D-structures of the respective scaffolds $\mathbf{A_{47}}$, $\mathbf{A_4}$, and $\mathbf{A_{36}}$, can retrospectively be rationalized by invoking the orientation of the three substituents relative to each other and to a plane defined by the aromatic core. Again, one may advise to choose only one out of the three libraries ($\mathbf{S_{47}}$, $\mathbf{S_4}$, $\mathbf{S_{36}}$) for a first round of library production. It is important to note that such recommendations can, at best, have the character of a first,

crude filtering tool that should be used merely to prioritize between the different sublibraries awaiting synthesis but never to exclude any of them from eventually being made. In our opinion, the envelope shape of a molecule—together with its size—constitutes the first, most basic, level in a hierarchy of molecular descriptors,[61,62] defining merely the playground for a wealth of secondary descriptors to refine the information with a more spatial view to the potential for interactions, such as polar surface area, hydrogen-bonding surface potential, surface charge distribution, or presence and location of specific pharmacophoric elements. For example, one may expect the compounds of sublibrary $\mathbf{S_{36}}$ derived from scaffold $\mathbf{A_{36}}$, while very similar to $\mathbf{S_{47}}$ and $\mathbf{S_4}$ in terms of envelope shape, to differ considerably in terms of, e.g., the surface charge distribution, due to the presence of a basic, at physiological pH probably protonated, nitrogen atom in $\mathbf{A_{36}}$. Thus, shape complementarity is *necessary, but alone not sufficient*, for a compound to productively interact with a target. As an extreme example, one might invoke the case of CH_4, BH_4^-, and NH_4^+, which all have the same shape (spherical), but entirely different electrostatic properties, making them prone to have different activities despite their identical shape and similar size. Nevertheless, it appears sensible to abide by the hierarchy and start with descriptors as basic as the envelope shape, simply because a molecule may well have the appropriate pharmacophoric elements, charge distribution, or other secondary features, yet will never get a chance to productively interact with the biological target, if it does not have the appropriate shape to fit the critical cavity.

So far, the analysis of the triangle plots and the NPR-based comparison of compound selections was performed by mere visual inspection. However, to avoid the potential subjective bias of such an approach, we sought to devise a general way to quantify the redundancy of the different sublibraries and their respective contributions to the overall shape diversity of the collection. Such an analysis could, for instance, be used to identify the minimum number of sublibraries required for maximum coverage of the triangle diagram, which would define the priorities for a first production cycle. Operating in a closed and well-defined coordinate system, the method is well suited for the application of cell-based diversity metrics. Standard programs might require some minor modifications to allow them to work on (trigonal) hyperprisms instead of hyperparallelepipeds. For illustration, we have subdivided the triangle area into 2500 triangles of equal size and calculated the corresponding membership counts for each of the 50 sublibraries of library (**C**). These were then used to compute pairwise Carbo indices as measures of interset similarity, as described in the Methodology section. As an illustration, Table 2 shows the Carbo indices (in %) obtained from pairwise comparisons between all sublibraries of library (**C**), $\mathbf{S_1}-\mathbf{S_{50}}$, based on membership counts in 2500 equally sized triangular bins. Gratifyingly, the numbers obtained from the quantitative analysis agree nicely with the impressions gained from visual inspection of the triangle graphs (see Chart 1, Supporting Information and Figure 9), with numbers superior to 70% indicating a high, in general also visually evident, similarity between two sublibraries.

This quantitative approach to determining shape similarity between different compound sets proves particularly useful in cases where mere visual inspection of the triangle diagrams cannot provide conclusive results. In the example shown in Figure 10, the degree of shape similarity between sublibrary $\mathbf{S_4}$ and sublibraries $\mathbf{S_{43}}$, $\mathbf{S_{45}}$, $\mathbf{S_{47}}$, $\mathbf{S_{49}}$ is difficult to assess based on the distribution patterns within the respective triangle diagrams. The quantitative analysis, however, reveals quite clearly that the shape space coverage of $\mathbf{S_4}$ is most similar to $\mathbf{S_{47}}$ and most dissimilar to that of $\mathbf{S_{45}}$, which would then suggest to prioritize $\mathbf{S_{45}}$ over $\mathbf{S_{47}}$ in the context of library construction, if $\mathbf{S_4}$ was already part of the existing set.

The quantitative method to assess shape similarity between compound sets also allows to check and further corroborate our conclusions previously drawn from optical inspection of the triangle graphs obtained from libraries (**A**), (**B**), (**C**) and MDDR (see Figures 6 and 7). Based on the graphs shown in Figure 6, the limited range of molecular shapes accessible to library (**A**) (or $\mathbf{S_1}$) appeared to be significantly improved only by increasing the number of central scaffolds, as in library (**C**), but not by increasing the number of peripheral substituents, as in library (**B**). Moreover, the plot obtained from library (**C**) seemed much more similar to the plot associated with the MDDR collection (see Figure 7) than that obtained from library (**B**), suggesting that the increase in shape space coverage achieved by maximizing the scaffold diversity, as in library (**C**), will increase the chances of addressing a range of different biological targets. Indeed, the Carbo indices obtained from pairwise comparison of libraries (**A**), (**B**), (**C**), MDDR, and sublibraries $\mathbf{S_2}-\mathbf{S_{50}}$ (see Table 3) show that the 125 000-member benzodiazepine library (**B**) remains very similar ($R_{AB} = 75\%$) to its 50-fold smaller counterpart, library (**A**), demonstrating that simply increasing the library size does not significantly improve the shape space coverage. On the other hand, the 109 850-member multiscaffold library (**C**) is highly dissimilar ($R_{AC} = 21\%$) to library (**A**). Furthermore, while the sheer *amount* of space covered by library (**A**), and, more so, by library (**B**), is not too much different from MDDR, they do fall into entirely different shape spaces, as revealed by a very low Carbo index of 22% and 15%, respectively. Conversely, a high similarity between the multiscaffold library (**C**) and MDDR of 73% is found, which confirms the impressions gained from visual inspection of the plots shown in Figures 6 and 7.

So far, we have exclusively considered "classical", three-point diversity libraries, in which the scaffold is relatively large compared to the size of the molecule, and shown that the molecular shape range covered by a given library is critically determined by the scaffold and the way it orients the substituents in three-dimensional space, rather than by the nature and/or the number of peripheral substituents. It could be argued, however, that the molecular shape is related to the scaffold only in the limit that the scaffold is large relative to the entire molecule, meaning that the influence of the scaffold on the overall shape will disappear with decreasing size. To investigate this, we took the extreme case of a library derived from carboxylic acids and amines, in which the "scaffold", i.e., the element remaining constant throughout the library, is reduced to an amide bond. To remain consistent with the preceding examples involving three-point diversity libraries, the same set of 50 diverse peripheral substituents, $\mathbf{B_1}-\mathbf{B_{50}}$, was used (see Figure 3), resulting in a set of 2500 amide compounds, termed $\mathbf{S_{51}}$. For

Table 2: Interset Similarity in Terms of Shape Space Coverage: Carbo Indices (in %) of Pairwise Comparisons between Libraries S_1-S_{50}[a]

S	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
01	-	15	19	0	2	2	1	8	2	1	15	12	9	14	5	4	29	46	12	4	2	21	5	4	34	59	20	46	6	7	8	23	7	61	39	2	26	17	7	59	23	11	2	3	4	24	3	16	1	6
02		-	79	14	48	42	22	67	21	34	42	79	25	59	46	37	26	41	42	37	26	26	28	41	33	33	43	41	53	62	73	46	29	33	41	23	50	56	33	7	78	32	58	31	67	71	18	77	40	20
03			-	12	49	37	27	57	17	31	39	73	19	53	40	31	22	40	37	29	18	22	26	32	29	38	39	38	48	55	66	42	24	38	39	19	47	52	28	10	75	26	49	24	61	71	15	78	34	15
04				-	47	65	70	33	82	68	24	25	53	28	65	37	12	6	30	53	46	8	79	52	15	2	21	12	68	29	28	19	51	4	13	84	20	26	75	0	6	33	43	62	27	7	89	4	68	70
05					-	84	51	50	40	78	15	57	20	36	44	26	8	10	25	32	24	5	52	42	15	7	21	13	65	42	58	18	23	9	11	46	20	28	44	1	24	17	73	34	52	24	44	28	65	21
06						-	60	58	55	88	24	54	30	40	57	38	11	12	33	43	35	8	69	52	19	7	25	17	75	48	56	23	35	9	15	62	24	34	61	1	23	26	73	47	51	23	62	22	76	35
07							-	33	63	61	29	34	46	34	59	38	16	11	35	49	45	12	67	48	18	5	25	15	64	35	33	23	48	8	18	65	25	31	62	1	14	33	44	51	36	15	69	11	64	52
08								-	46	55	58	80	43	70	70	68	31	35	64	65	53	29	50	73	40	23	55	44	72	81	82	50	48	27	42	47	54	65	53	4	55	49	77	60	81	55	41	48	70	35
09									-	57	38	34	69	40	72	53	21	13	45	68	64	14	72	68	23	6	31	19	67	41	36	28	62	8	21	86	30	38	69	1	12	46	50	81	40	14	85	9	72	73
10										-	25	47	34	38	58	41	13	10	35	46	40	8	73	55	20	5	25	17	74	46	51	23	37	8	15	66	24	33	65	1	17	28	69	48	47	17	66	15	79	38
11											-	55	60	74	64	73	58	53	76	69	61	56	48	62	56	35	74	59	52	72	56	69	66	40	64	32	71	76	50	9	49	70	32	58	57	51	33	30	36	45
12												-	38	71	61	55	33	40	58	53	41	29	42	61	41	28	54	46	66	77	83	52	43	33	43	36	55	67	44	7	65	43	77	48	82	63	29	60	59	30
13													-	58	67	64	52	34	65	75	73	47	63	68	48	18	58	43	54	50	35	57	76	23	48	60	55	60	62	7	21	73	35	78	43	27	60	12	49	69
14														-	68	73	56	51	75	70	61	52	52	67	57	33	71	57	65	80	72	71	65	37	59	36	70	78	53	9	55	69	55	62	75	58	34	43	50	42
15															-	77	35	30	68	81	74	31	74	81	41	16	59	41	79	72	64	53	72	22	43	70	53	68	75	3	38	62	61	81	69	38	71	28	74	63
16																-	45	32	79	83	81	40	60	79	48	16	67	48	64	72	60	63	74	22	49	44	59	74	60	3	32	68	50	75	66	35	46	21	57	47
17																	-	63	63	46	39	76	34	39	73	46	69	63	31	44	29	69	58	48	64	17	67	57	36	21	32	70	15	35	32	41	18	16	16	37
18																		-	51	31	21	66	24	28	66	64	62	67	31	44	34	63	38	62	67	11	67	53	27	29	51	50	17	22	33	56	11	35	13	26
19																			-	78	70	61	55	72	61	29	75	59	60	73	59	72	74	36	59	38	71	75	57	9	41	77	44	65	64	46	39	26	49	48
20																				-	85	41	68	81	46	15	63	44	69	67	55	60	80	21	49	59	57	69	69	3	31	73	51	82	63	34	60	20	64	59
21																					-	31	63	76	38	9	55	36	58	56	44	50	78	15	41	52	47	63	62	2	19	68	40	81	53	22	55	11	56	51
22																						-	29	31	71	41	71	62	28	44	25	69	53	44	62	11	68	53	33	15	35	65	9	27	29	46	14	17	12	33
23																							-	65	36	13	48	33	75	49	44	43	68	17	35	75	43	51	82	3	21	55	50	71	45	24	81	14	69	66
24																								-	43	15	56	40	72	72	65	52	69	19	41	64	51	66	67	3	32	61	66	80	69	34	60	22	75	56
25																									-	45	69	67	39	50	38	68	54	52	63	21	67	59	38	25	36	62	26	37	38	44	21	21	26	34
26																										-	37	55	18	25	23	44	19	66	57	6	48	37	16	42	44	29	10	10	18	47	7	33	7	14
27																											-	66	51	66	52	74	67	44	67	27	74	74	48	14	47	74	33	51	55	52	28	29	34	42
28																												-	37	53	42	66	48	60	66	17	67	61	36	36	47	58	24	34	41	52	18	31	22	31
29																													-	70	68	48	58	23	36	70	49	60	72	4	40	51	72	67	67	42	69	32	82	57
30																														-	80	64	56	30	51	41	64	74	52	4	55	61	65	59	81	55	35	40	59	39
31																															-	49	41	26	38	37	53	66	47	4	59	42	74	50	82	58	31	54	60	30
32																																-	65	46	69	23	75	72	47	16	50	75	31	46	50	54	25	35	29	41
33																																	-	25	54	56	61	66	66	5	26	79	35	74	49	33	58	16	50	61
34																																		-	58	8	52	40	20	48	43	33	12	16	23	47	10	34	10	17
35																																			-	18	70	64	38	28	52	62	19	37	37	58	21	35	21	33
36																																				-	26	33	74	1	14	37	54	72	38	13	87	10	74	75
37																																					-	72	46	18	57	71	32	45	51	62	27	40	31	43
38																																						-	53	11	56	69	45	61	69	59	33	42	43	43
39																																							-	5	27	55	49	69	48	29	79	18	67	67
40																																								-	11	9	1	2	2	13	2	7	1	5
41																																									-	33	33	22	52	79	11	80	22	18
42																																										-	31	63	50	39	41	17	39	54
43																																											-	54	75	33	46	34	80	30
44																																												-	56	25	70	15	67	67
45																																													-	53	32	46	61	33
46																																														-	12	73	21	21
47																																															-	7	70	71
48																																																-	17	9
49																																																	-	48
50																																																		-

[a] (See Table 1, Figures 3 and 4, Chart 1, Supporting Information), based on membership counts in 2500 equally sized triangular bins (see Methodology Section).

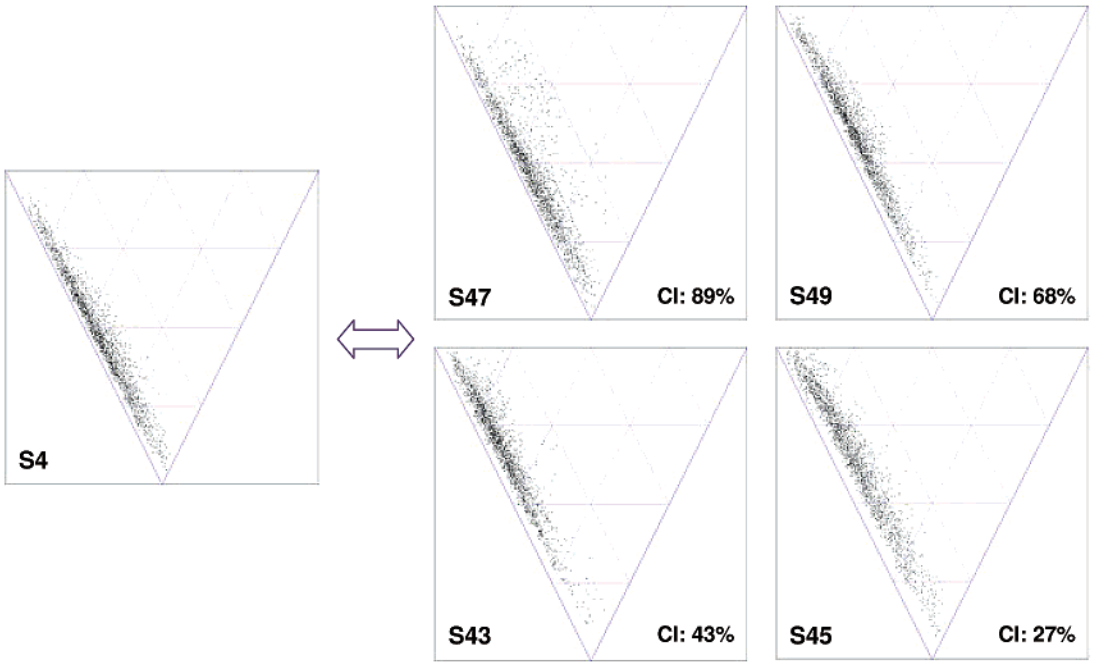


Figure 10. Quantitative interset similarity analysis based on shape space coverage: beyond the subjective bias of visual inspection (CI = Carbo index).

Table 3. Carbo Indices (in %) of Pairwise Comparisons between Libraries (A) = S_1, (B), (C), MDDR and S_2-S_{50}[a]

	A	B	C	MDDR	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11	S12	S13	S14
A		75	21	22	100	15	19	0	2	2	1	8	2	1	15	12	9	14
B			15	15	75	6	8	0	0	1	0	3	1	0	8	6	7	9
C				73	21	64	58	61	57	66	60	81	70	65	74	77	73	82
MDDR					22	71	79	31	68	61	46	65	34	53	45	77	34	62
	S15	S16	S17	S18	S19	S20	S21	S22	S23	S24	S25	S26	S27	S28	S29	S30	S31	S32
A	5	4	29	46	12	4	2	21	5	4	34	59	20	46	6	7	8	23
B	2	2	25	35	8	2	1	16	3	2	31	48	17	43	3	3	3	16
C	87	79	56	53	80	83	73	51	78	83	60	37	73	60	86	82	76	71
MDDR	53	42	32	43	48	44	33	31	47	48	39	37	46	44	68	63	72	48
	S33	S34	S35	S36	S37	S38	S39	S40	S41	S42	S43	S44	S45	S46	S47	S48	S49	S50
A	7	61	39	2	26	17	7	59	23	11	2	3	4	24	3	16	1	6
B	5	59	31	1	18	11	5	79	9	8	0	1	1	12	2	5	0	5
C	77	42	62	68	72	80	78	15	56	73	71	79	77	59	67	45	76	65
MDDR	38	39	44	37	52	57	47	14	63	38	65	40	72	67	33	62	54	29

[a] (See Table 1, Figures 3 and 4, Chart 1, Supporting Information), based on membership counts in 2500 equally sized triangular bins (see Methodology Section).

the purpose of comparison, a series of analogous 2500-member two-point diversity libraries, $\mathbf{S_{52}}-\mathbf{S_{65}}$, around related "minimal" scaffolds was constructed as described in the Methodology section (see Table 1, Figures 3 and 5), in which the substituents are equally separated by only two, in a few cases three, atoms. The computational treatment through to the generation of triangular graph plots (see Chart 2, Supporting Information) and computation of pairwise Carbo similarity indices (see Table 4) followed the procedures outlined in the Methodology section. Somewhat surprisingly, both the graphic and the numeric results clearly indicate that even these very small scaffolds exert a highly discriminating effect on the molecular shape distribution of the respective libraries. The shape range associated with the amide library $\mathbf{S_{51}}$ appears most similar to the ester library $\mathbf{S_{52}}$ (but *not* to the sulfonamide library $\mathbf{S_{53}}$), the (*E*)-olefin library $\mathbf{S_{54}}$ (but *not* its (*Z*)-olefin counterpart $\mathbf{S_{55}}$), the azo-compound library $\mathbf{S_{56}}$, and finally, to a lesser extent, also the oxadiazole and thiazoline libraries $\mathbf{S_{58}}$ and $\mathbf{S_{59}}$ (but *not* the closely related imidazoline library $\mathbf{S_{60}}$, which appears more akin to the sulfonamide library $\mathbf{S_{53}}$). Of interest, both (*E*)-double bonds,[63] azo-groups,[64] oxadiazoles,[65] and thiazolines[66] have previously been reported to behave as bioisosteric replacements for the amide bond. However, the fact that the same is true for imidazolines,[67] which, at least in our analysis, are much less similar to amide compounds in terms of molecular shape, is a reminder that shape similarity is not always necessary (or sufficient) for bioisosterism and may, in some cases, be set off (or complemented) by, e.g., electronic factors. Finally, given the high degree of shape similarity between the amide and the (*E*)-alkene libraries, the low similarity between the amide library $\mathbf{S_{51}}$ and the small carbocyclic libraries with vicinal orientation of the substituents, $\mathbf{S_{62}}-\mathbf{S_{65}}$, is rather surprising. Although, as one would predict, the two libraries with *trans*-oriented substituents, $\mathbf{S_{63}}$ and $\mathbf{S_{65}}$, are still more similar to $\mathbf{S_{51}}$ than the corresponding *cis*-versions, $\mathbf{S_{62}}$ and

Table 4. Carbo Indices (in %) of Pairwise Comparisons between Libraries $S_{51}-S_{65}$[a]

S	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65
51		74	37	80	28	48	82	71	73	35	78	18	32	25	51
52			60	73	46	76	65	67	88	53	75	31	42	44	78
53				40	73	71	29	48	52	79	51	63	70	68	71
54					30	58	84	82	77	37	74	16	30	28	56
55						57	20	37	38	73	37	70	71	77	58
56							41	62	76	71	71	37	51	54	76
57								77	61	29	65	9	23	17	41
58									68	46	76	23	37	32	57
59										43	82	23	34	36	72
60											50	54	69	64	65
61												22	40	31	63
62													66	63	44
63														67	53
64															55
65															

[a] (See Table 1, Figures 3 and 5, Chart 2, Supporting Information), based on membership counts in 2500 equally sized triangular bins (see Methodology Section).

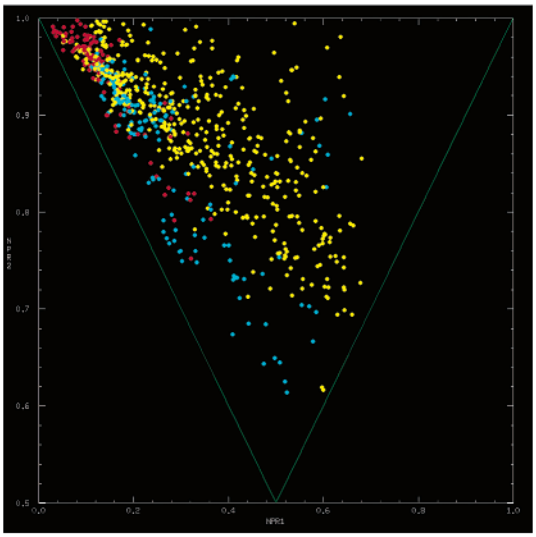
S_{64}, the absolute values of all four Carbo indices remain far behind the one calculated for the (*E*)-alkene library, indicating that very subtle changes in the scaffold geometry can indeed have a drastic impact on the overall molecular shape distribution, and this holds even if the scaffold is very small compared to the rest of the molecule.

All data presented so far underline the paramount importance of the scaffold in defining molecular shape and strongly advise to include as many different scaffolds as possible when assembling a compound collection aimed at maximum shape diversity. The link between shape diversity and biological activity, however, has remained less well defined, despite the observation that the multiscaffold library (**C**), unlike the single-scaffold library (**B**), produces a shape distribution pattern very similar to that associated with MDDR, i.e., a compound collection covering more than 500 types of biological activities (see Figures 6 and 7 and Table 3). Looking at space-filling representations of cocrystal structures of small molecule inhibitors and enzyme targets, one would tend to accept that molecular shape has something to do with biological activity, since it is intuitively clear that a compound will only modulate the activity of a biological target, if its 3D-shape can match the appropriate cavities presented by the biological counterpart. On the other hand, as pointed out before (see Figure 9 and Discussion), shape complementarity is necessary, but alone not sufficient for a compound to productively interact with a target, meaning that molecules with similar shapes will not necessarily produce similar biological activities. Conversely, one might ask whether molecules with different shapes will display different biological activities. To investigate this, we randomly extracted a number of subsets from MDDR consisting of at least 250 compounds (average 700 compounds/set) reported to interact with molecularly defined biological targets and analyzed them with respect to their molecular shape similarity. The resulting shape triangle diagrams shown in Chart 3, Supporting Information, as well as the corresponding Carbo indices (in %) reported in Table 5 indicate that the interset similarity is generally very low, with an average value over all pairwise comparisons of 22%, and a range from 0% to 73%, suggesting that there may indeed be a trend toward different biological activities being associated

Table 5. Carbo Indices (in %) of Pairwise Comparisons between MDDR Compound Sets Grouped by Biological Activities[a]

MDL-No	Activity Class	#Cpds	31261	06235	06240	06245	07701	07710	09249	12455	12457	31410	31432	31500	37110	42713	43210	52500	54112	65000	68210	71522	78348	78418	78429	78454
31261	*α1 ant*	444	-	52	57	36	53	50	45	33	5	13	3	45	34	4	34	13	42	4	15	11	40	10	24	7
06235	*5-HT1A ag*	833		-	55	38	73	62	34	34	7	9	0	45	25	1	28	18	42	0	14	9	52	10	21	7
06240	*5-HT1A ant*	341			-	43	48	48	42	31	4	16	7	42	32	4	33	18	38	1	19	10	39	12	24	7
06245	*SSRIs*	286				-	33	42	40	30	11	10	7	37	29	12	27	20	32	2	15	7	25	15	21	14
07701	*D2 ant*	389					-	57	31	29	7	9	0	44	25	1	21	19	36	0	15	10	51	9	24	7
07710	*D4 ant*	471						-	41	31	5	11	0	55	33	2	30	18	40	0	20	15	41	6	23	5
09249	*M1 ag*	929							-	34	10	18	12	50	38	12	42	24	37	10	22	14	30	20	25	14
12455	*NMDA-R ant*	1362								-	44	21	10	33	32	10	32	31	34	5	29	14	31	35	26	24
12457	*AMPA-R ant*	341									-	11	5	1	10	4	15	32	19	0	27	6	9	27	14	29
31410	*ACE inh*	517										-	29	25	32	20	24	24	18	12	14	8	13	26	27	13
31432	*AT1 ant*	517											-	15	29	19	24	25	10	15	12	6	4	28	19	8
31500	*Ca-Channel blocker*	1479												-	44	17	41	18	29	8	14	16	37	17	25	8
37110	*Thombin inh.*	741													-	17	35	25	30	11	22	14	26	31	34	14
42713	*CCKB ant*	296														-	17	14	5	13	9	2	3	21	16	11
43210	*Aldose Red. inh*	902															-	30	31	16	28	14	26	28	33	16
52500	*HMG-CoA Red. Inh*	1031																-	33	6	59	11	17	46	39	42
54112	*H/K-ATPase inh*	644																	-	5	25	12	42	26	27	24
65000	*Antibiotic Macrolide*	578																		-	2	6	4	11	7	0
68210	*Antibiotic Quinolone*	1046																			-	8	16	37	27	23
71522	*Rev. Transcript. inh*	379																				-	10	13	16	6
78348	*PLA2 inh*	633																					-	11	19	5
78418	*PDE IV inh*	1016																						-	31	30
78429	*Farnesyl Tranfer. inh*	555																							-	25
78454	*COX-2 inh*	377																								-

[a] (See Chart 3, Supporting Information), based on membership counts in 2500 equally sized triangular bins (see Methodology Section).



	Class 1	Class 2	Class 3	All (31500)
Class 1 (Nifedipine-like)	-	20	19	62
Class 2 (Diltiazem-like)		-	16	51
Class 3 (Verapamil-like)			-	53
All (MDDR-class 31500)				-

pair-wise Carbo indices (in %)

Figure 11. Comparison of shape similarities associated with different chemical classes of L-type, voltage-gated Ca^{2+}-channel blockers belonging to MDDR activity class #31500.

with different molecular shapes. More importantly, however, when the comparison is restricted to compound sets interacting with closely related biological targets, such as biogenic amine GPCRs (31261, 06235, 06240, 07701, 07710, 09249), a markedly higher average interset similarity of 50% is observed, along with (as one might expect) high interset similarities between, e.g., D_2- and D_4-antagonists (57%), or 5-HT_{1A}-agonists and antagonists (55%). Somewhat unexpectedly, the highest interset similarities result from comparisons of 5-HT_{1A}-agonists with D_2-antagonists (73%) and D_4-antagonists (62%). Notably, this high average similarity does not universally apply to all compound sets interacting with GPCRs. There seems to be a strong discrimination between GPCR subclasses based on the nature of their endogenous ligands (biogenic amines versus peptides), as illustrated by the strikingly low similarities (0−12%) between the sets containing adrenergic, serotoninergic, dopaminergic, and muscarinic ligands (31261, 06235, 06240, 07701, 07710, 09249) versus those consisting of AT_1- and CCK_B-antagonists (31432, 42713). While the data presented in Table 5 indeed lend support to the notion that different shapes are often correlated with different biological activities, one must be precautious not to generalize, based on several considerations: The biological target and the overall shapes of its binding sites represent a problem which can, in principle, be solved in many different ways. Thus molecules may fill out the entire pocket (in which case they would probably tend to be more elongated) or just parts of the pocket (in which case they might be more compact, thus spherical/discoid). A second layer of complexity is added by the fact that many biological targets contain several different sites prone to interact with chemical agents, a prime example being the kinases, which, in addition to the main ATP- and substrate binding domains, often contain additional "allosteric" modulation sites as well as regulatory protein−protein interaction sites such as PH- or SH2-domains. If the final objective is inhibition-of-function of a given kinase, irrespective of the mechanism, then this may well be achieved by different chemical series with different molecular shapes. The L-type, voltage-gated Ca^{2+}-channel constitutes another classical example of this concept. Three main chemical classes of L-type Ca^{2+}-channel blockers have been identified and shown to interact at three spatially distinct sites of the channel, namely the dihydropyridines (e.g. nifedipine), the benzothiazepines (e.g. diltiazem), and the phenylalkylamines (e.g. verapamil).[68] With this in mind, we decided to reanalyze the corresponding MDDR activity class containing all types of Ca-channel blockers (#31500, see also Table 5), focusing this time on the distribution patterns generated by the three different chemical classes mentioned before. As expected, both the triangle plot and the quantitative analysis reveal the three classes of Ca^{2+}-channel blockers to fall into clearly distinguishable shape spaces, with low interclass similarity indices of ≤20% (see Figure 11), thus constituting an experimental example of compound sets with dissimilar molecular shape distribution, yet similar biological activity. Taken together, the results presented above corroborate the notion that molecular shape is correlated with biological activity and that a high degree of shape (hence scaffold) diversity in screening collections will increase the odds of addressing a broad range of biological targets. Clearly, one can cite numerous exceptions, in which chemical series with different shapes produce similar biological activities, either by interacting in a different way with the same site or by addressing different sites on the same biological target. It is noteworthy, however, that these exceptions do not contradict the basic thesis of striving for maximum shape diversity in primary screening collections, because it is highly desirable to identify several distinct chemical series active on a given target, to preempt potential downstream issues often encountered with one chemical family, such as toxicity or poor pharmacokinetics.

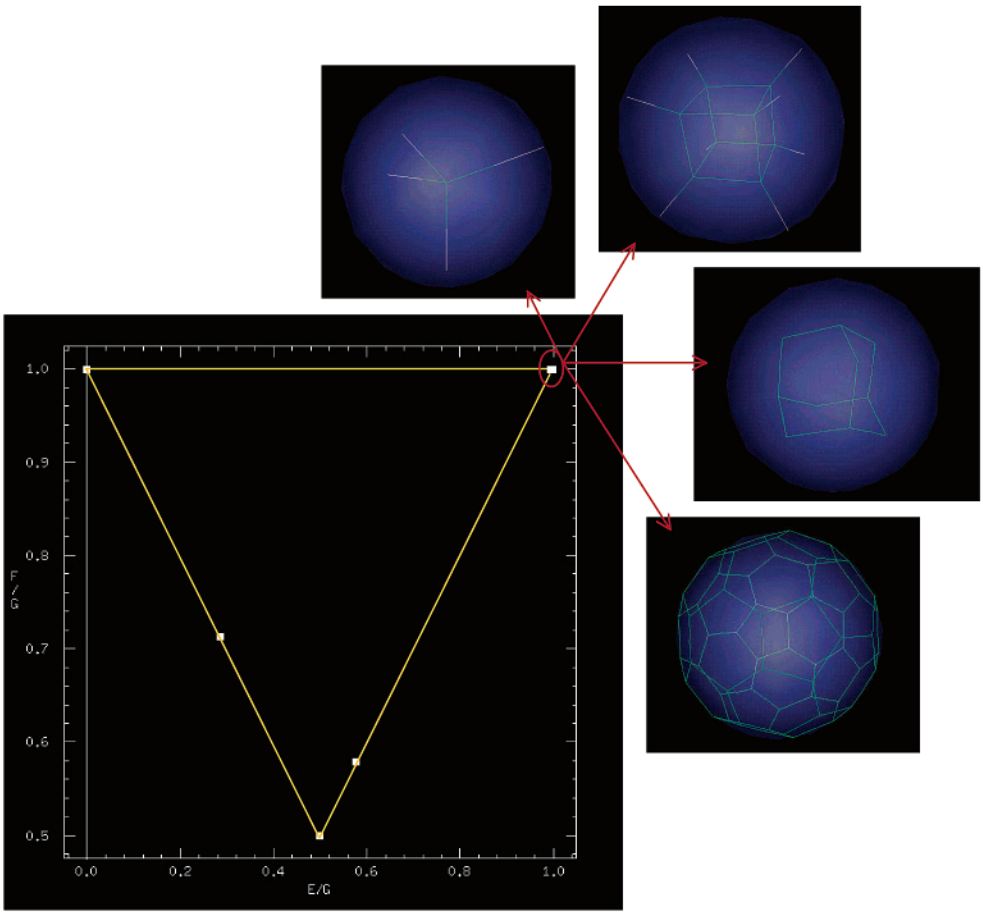


Figure 12. Limitations of the envelope shape analysis: degenerate situations.

The envelope shape analysis offers a rapid way to assess and compare compound collections for their biorelevant diversity, but, as every method, it has some obvious limitations, such as the fact that it appears to disregard any information related to the situation on the inside of the molecular envelope. Thus, molecules with the same envelope shape will be deemed identical, even though they may substantially differ with respect to, e.g., the extent to which the molecular surface is rugged on the inside of the envelope or the degree to which the molecule fills the space defined by its envelope. As an illustration of this point, Figure 12 shows the extreme example of four molecules as different as methane, cubane, adamantane, and fullerene, which cannot be distinguished based on their envelope shape. However, further studies in our lab have since shown that such "degenerate" situations are extremely rare and restricted only to highly symmetrical, unsubstituted molecules. For example, it was found that, quite surprisingly, a monofluoro substitution is already sufficient to break the degeneration, with fluoromethane, fluorocubane, and fluoroadamantane being significantly spread apart in the shape diagram, probably because the symmetry disturbance caused by the fluorine atom decreases with increasing absolute molecular volume. Still more surprisingly, however, it was noted that even tetrakis-substitution with four *identical* R-groups (except those having local C_∞-symmetry) leads to a rupture in envelope shape degeneration. The effect is so significant that it was used to tackle one of the classical problems of dissimilarity-based library comparison, namely the separation of a cubane from an adamantane library with four tetragonally oriented substitution positions and four identical sets of substituents.[69] The results of this study will be reported in a subsequent article.[70]

CONCLUSIONS

We have developed a computational method to rapidly assess and visualize the diversity in molecular shape associated with a given compound set, making use of normalized ratios of principal moments of inertia plotted into two-dimensional triangular graphs (NPR-analysis). Using this approach to compare the shape space covered by different compound collections, such as combinatorial libraries or sets of bioactive compounds, we have shown that any given scaffold leads to an idiosyncratic localized distribution pattern in the triangle graph pointing to a limited range of accessible molecular shapes. Increasing the number of peripheral substituents as well as allowing for multiple conformations per compound does not significantly improve the shape space coverage. On the other hand, combining several small libraries around distinct chemical scaffolds was shown to produce a compound set with a high degree of molecular shape diversity, underlining the pivotal role of the scaffold and its 3D-geometry in defining the molecular shape range accessible to the members of a given combinatorial library. We have further developed a computational method to quantify interset similarity in terms of shape space coverage, which allows to calculate the shape redundancy between the different sublibraries of a given compound collection. Useful

applications include the prioritization between different sublibraries to be synthesized or purchased based on their relative contribution to the overall shape diversity of an existing collection or the selection of appropriate focused screening sets based on a similar shape space coverage compared to compounds known to interact with a given biological target. Maximum shape space coverage was further shown to be correlated with, and probably necessary for, broad biological activity by an NPR-analysis involving collections of known bioactive compounds, such as MDDR and the GOLD-set. A comparison of the molecular shape distribution patterns associated with different MDDR subsets of known biological activities further corroborates the intuitive notion that molecular shape is intimately linked to biological activity and that a high degree of shape (hence scaffold) diversity in screening collections will increase the odds of addressing a broad range of biological targets.

ACKNOWLEDGMENT

We would like to thank our colleagues Agnès Bombrun (SPRI, Geneva) and Xuliang Jiang (SRBI, Rockland, U.S.A.) for critically revising this manuscript; Nabil El Tayar for Figure 2, and Prof. Hugo Kubinyi and the anonymous reviewers for valuable and encouraging comments.

Supporting Information Available: Charts showing the individual shape triangle diagrams obtained by plotting the envelope shapes of the respective compounds belonging to libraries $\mathbf{S_1}-\mathbf{S_{50}}$ (Chart 1, see Table 2 in the main text), $\mathbf{S_{51}}-\mathbf{S_{65}}$ (Chart 2, see Table 4 in the main text), and the MDDR biological activity subsets (Chart 3, see Table 5 in the main text), using the methods outlined in the Methodology section. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Balkenhohl, F.; von dem Bussche-Huennefeld, C.; Lansky, A.; Zechel, C. Combinatorial Synthesis of Small Organic Molecules. *Angew. Chem., Int. Ed. Engl.* **1996**, *35*, 2288−2337.
(2) Thompson, L. A.; Ellman J. A. Synthesis and Applications of Small Molecule Libraries. *Chem. Rev.* **1996**, 555−600.
(3) Gordon, E. M.; Gallop, M. A.; Patel, D. V. Strategy and Tactics in Combinatorial Organic Synthesis. Applications to Drug Discovery. *Acc. Chem. Res.* **1996**, *29*, 144−154.
(4) Katritzky, A. R.; Kiely, J. S.; Hebert, N.; Chassaing, C. Definition of Templates within Combinatorial Libraries. *J. Comb. Chem.* **2000**, *2*, 2−5.
(5) Weber, L. High-diversity combinatorial libraries. *Curr. Opin. Chem. Biol.* **2000**, *4*, 295−302.
(6) Maggiora, G. M.; Shanmugasundaram, V. On the similarity of chemistry spaces. *Abstr. Pap. − Am. Chem. Soc.* **2001**, *221st*, COMP-077.
(7) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.
(8) Katritzky, A. R.; Gordeeva, E. V. Traditional topological indexes vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 835−857.
(9) Matter, H. Selecting Optimally Diverse Compounds from Structure Databases: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.
(10) Matter, H. A validation study of molecular descriptors for the rational design of peptide libraries. *J. Pept. Res.* **1998**, *52*, 305−314.
(11) Palm, K.; Luthman, K.; Ungell, A. L.; Strandlund, G.; Beigi, F.; Lundahl, P.; Artursson, P. Evaluation of dynamic polar molecular surface area as predictor of drug absorption: comparison with other computational and experimental predictors. *J. Med. Chem.* **1998**, *41*, 5382−5392.
(12) Bursi, R.; Dao, T.; van Wijk, T.; de Gooyer, M.; Kellenbach, E.; Verwer, P. Comparative spectra analysis (CoSA): spectra as three-dimensional molecular descriptors for the prediction of biological activities. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 861−867.
(13) Xue, L.; Godden, J.; Gao, H.; Bajorath, J. Identification of a preferred set of molecular descriptors for compound classification based on principal component analysis. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 699−704.
(14) Xue, L.; Bajorath, J. Molecular descriptors for effective classification of biologically active compounds based on principal component analysis identified by a genetic algorithm. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 801−809.
(15) Xue, L.; Godden, J. W.; Bajorath, J. Evaluation of descriptors and mini-fingerprints for the identification of molecules with similar activity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1227−1234.
(16) Labute, P. A widely applicable set of descriptors. *J. Mol. Graphics Modell.* **2000**, *18*, 464−477.
(17) Brown, R. D.; Martin, Y. C. Use of Structure−Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.
(18) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand−receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.
(19) Martin, Y. C.; Bures, M. G.; Brown, R. D. Validated descriptors for diversity measurements and optimization. *Pharm. Pharmacol. Commun.* **1998**, *4*, 147−152.
(20) Basak, S. C.; Harriss, D. K.; Magnuson, V. R. Comparative study of lipophilicity versus topological molecular descriptors in biological correlations. *J. Pharm. Sci.* **1984**, *73*, 429−437.
(21) Randic, M. Orthogonal molecular descriptors. *New J. Chem.* **1991**, *15*, 517−525.
(22) Randic, M. Generalized molecular descriptors. *J. Math. Chem.* **1991**, *7*, 155−168.
(23) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(24) Bayada, D. M.; Hamersma, H.; van Geerestein, V. J. Molecular diversity and representativity in chemical databases. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1−10.
(25) Pearlman, R. S.; Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.
(26) Stanton, D. T. Evaluation and Use of BCUT Descriptors in QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 11−20.
(27) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring Diversity: Experimental Design of Combinatorial Libraries for Drug Discovery. *J. Med. Chem.* **1995**, *38*, 1431−1436.
(28) Pickett, S. D.; Luttmann, C.; Guerin, V.; Laoui, A.; James, E. DIVSEL and COMPLIB − strategies for the design and comparison of combinatorial libraries using pharmacophoric descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 144−150.
(29) Xue, L.; Bajorath, J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Comb. Chem. High Throughput Screen.* **2000**, *3*, 363−372.
(30) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(31) Walters, W. P.; Ajay; Murcko, M. A. Recognizing molecules with drug-like properties. *Curr. Opin. Chem. Biol.* **1999**, *3*, 384−387.
(32) Meyer, A. Y. The Size of Molecules. *Chem. Soc. Rev.* **1986**, *15*, 449−474.
(33) Kier, L. B.; Hall, L. H. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure−Property Modeling. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, U.S.A., 1991; Vol. 2, pp 367−422.
(34) Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure−Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323−2329.
(35) Stouch, T. R.; Jurs, P. C. A Simple Method for the Representation, Quantification, and Comparison of the Volumes and Shapes of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 4−12.
(36) catShape 4.5, part of the Catalyst package; Accelrys Inc.: San Diego, CA, 2001.
(37) Broto, P.; Moreau, G.; Vandycke, C. Molecular structures: perception, autocorrelation descriptor and SAR studies. *Eur. J. Med. Chem. − Chim. Ther.* **1984**, *19*, 61−65, 66−70, 71−78, 79−84.
(38) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, 2000; pp 352, 390−395.
(39) E.g., 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*, 3919−3933. Good, A. C.; Richards, W. G. Rapid Evaluation of Shape Similarity Using Gaussian Functions. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 112−116. Meyer, A. Y.; Richards,

W. G. Similarity of molecular shape. *J. Comput.-Aided Mol. Design* **1991**, *5*, 426−439.
(40) E.g., Van de Waterbeemd, H.; Camenisch, G.; Folkers, G.; Raevsky, O. A. Estimation of Caco-2 cell permeability using calculated molecular descriptors. *QSAR* **1996**, *15*, 480−490. Ozaki, Y.; Shinoda, Y.; Ichihashi, M.; Kondow, T. Spheroidal vibrations of Ar-cluster isomers analyzed by molecular dynamics calculation. *Surf. Rev. Lett.* **1996**, *3*, 479−482. Maiti, S.; Jaman, A. I.; Datta, A.; Nandi, R. N. Microwave spectrum of 2,3-difluorobenzonitrile. *J. Mol. Spectrosc.* **1990**, *140*, 416−418. Schultz, G.; Hargittai, I.; Friedman, P. The molecular structure of isothiazole from electron diffraction and ab initio calculations. *J. Mol. Struct.* **1988**, *176*, 61−69. Richards, R. J.; Davis, R. W.; Gerry, M. C. L. The microwave spectrum and structure of chlorine thiocyanate. *J. Chem. Soc., Chem. Commun.* **1980**, 915−916. Chadwick, D.; Legon, A. C.; Millen, D. J. Microwave spectrum and ring planarity of cyclopent-2-en-1-one. *J. Chem. Soc. D* **1969**, 1130−1131. Downs, A. J.; Schmutzler, R. Stereochemistry of fluorophosphoranes. I. The vibrational spectrum and molecular structure of methyltetrafluorophosphorane, CH_3PF_4. *Spectrochim. Acta* **1965**, *21*, 1927−1939. Loos, K. R.; Lord, R. C. Vibrational spectrum and barrier to internal rotation for CF_3CFO. *Spectrochim. Acta* **1965**, *21*, 119−125.
(41) Cerius^2 Modelling Environment; Accelrys Inc.: San Diego, CA.
(42) Molecular Operating Environment; Chemical Computing Group: Montreal, Canada.
(43) Sybyl. Tripos Inc.: St. Louis, MO.
(44) Tsar (Tools for Structure Activity Relationships). Accelrys Inc.: San Diego, CA.
(45) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic generation of 3D atomic coordinates for organic molecules. *Tetrahedron Comput. Methodol.* **1990**, *3*, 537−547.
(46) Sadowski, J.; Rudolph, C.; Gasteiger, J. The generation of 3D models of host−guest complexes. *Anal. Chim. Acta* **1992**, *265*, 233−241.
(47) Dolle, R. E. Comprehensive survey of combinatorial libraries with undisclosed biological activity: 1992−1997. *Mol. Diversity* **1998**, *4*, 233−256.
(48) Dolle, R. E.; Nelson, K. H., Jr. Comprehensive survey of combinatorial library synthesis: 1998. *J. Comb. Chem.* **1999**, *1*, 235−282.
(49) Dolle, R. E. Comprehensive survey of combinatorial library synthesis: 1999. *J. Comb. Chem.* **2000**, *2*, 383−433.
(50) Dolle, R. E. Comprehensive survey of combinatorial library synthesis: 2000. *J. Comb. Chem.* **2001**, *3*, 477−517.
(51) Carbo, R.; Leyda, L.; Arnau, M. An electron density measure of the similarity between two compounds. *Int. J. Quantum Chem.* **1980**, *17*, 1185−1189.
(52) Hodgkin, E. E.; Richards, W. G. Molecular similarity based on electrostatic potential and electric field. *Int. J. Quantum Chem. Quantum Biol. Symp.* **1987**, *14*, 105−110.
(53) Asp 3.23 Reference Guide; Oxford Molecular: Oxford, UK, 2000; pp 2-2−2-3.
(54) Note that "fraction of compounds" will only give values different from those obtained from "number of compounds", if the Hodgkin formula is employed and the two libraries differ in their respective total numbers of compounds.
(55) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748. The data set is available online at http://www.ccdc.cam.ac.uk/prods/gold/value.html.
(56) Allen, F. H.; Bellard, S.; Brice, M. D.; Cartwright, B. A.; Doubleday: A.; Higgs, H.; Hummelink, T.; Hummelink-Peters, B. G.; Kennard, O.; Motherwell, W. D. S.; S.Rodgers, J. R.; Watson D. G. The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information. *Acta Crystallogr.* **1979**, *B35*, 2331−2339.
(57) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000−1008.
(58) Sadowski, J.; Gasteiger, J. From atoms and bonds to three-dimensional atomic coordinates: automatic model builders. *Chem. Rev.* **1993**, *93*, 2567−2581.
(59) E.g., Nicklaus, M. C.; Wang, S.; Driscoll, J. S.; Milne, G. W. A. Conformational Changes Of Small Molecules Binding To Proteins. *Bioorg. Med. Chem.* **1995**, *3*, 411−428.
(60) catConf 4.5, part of the Catalyst package; Accelrys Inc.: San Diego CA, 2001.
(61) Gasteiger, J.; Kleinoder, T.; Sadowski, J.; Wagener, M.; Hemmer, M. C. A hierarchy of structure representation. *Abstr. Pap. − Am. Chem. Soc.* **2002**, 223rd, CINF-002.
(62) Brüstle, M.; Beck, B.; Schindler, T.; King, W.; Mitchell, T.; Clark, T. Descriptors, Physical Properties, and Drug-Likeness. *J. Med. Chem.* **2002**, *45*, 3345−3355.
(63) E.g., Andres, C. J.; Macdonald, T. L.; Ocain, T. D.; Longhi, D. Conformationally defined analogs of prolylamides. Trans-Prolyl peptidomimetics. *J. Org. Chem.* **1993**, *58*, 6609−6613. Baginski, M.; Piela, L.; Skolnick, J. The ethylene group as a peptide bond mimicking unit: a theoretical conformational analysis. *J. Comput. Chem.* **1993**, *14*, 471−477. Shue, Y. K.; Tufano, M. D.; Carrera, G. M., Jr.; Kopecka, H.; Kuyper, S. L.; Holladay, M. W.; Lin, C. W.; Witte, D. G.; Miller, T. R. Double bond isosteres of the peptide bond: synthesis and biological activity of cholecystokinin (CCK) C-terminal hexapeptide analogs. *Bioorg. Med. Chem.* **1993**, *1*, 161−171.
(64) E.g., Kagechika, H.; Kawachi, E.; Hashimoto, Y.; Shudo, K. Differentiation inducers of human promyelocytic leukemia cells HL-60. Azobenzenecarboxylic acids and stilbenecarboxylic acids. *Chem. Pharm. Bull.* **1985**, *33*, 5597−5600. Kagechika, H.; Himi, T.; Namikawa, K.; Kawachi, E.; Hashimoto, Y.; Shudo, K. Retinobenzoic acids. 3. Structure−activity relationships of retinoidal azobenzene-4-carboxylic acids and stilbene-4-carboxylic acids. *J. Med. Chem.* **1989**, *32*, 1098−1108.
(65) E.g., Adelstein, G. W. Antiarrhythmic agents. Synthesis and biological activity of some tetrazole and oxadiazole analogs of 4-dialkylamino-2,2-diarylbutyramides. *J. Med. Chem.* **1973**, *16*, 309−312. Harfenist, M.; Heuser, D. J.; Joyner, C. T.; Batchelor, J. F.; White, H. L. Selective Inhibitors of Monoamine Oxidase. 3. Structure−Activity Relationship of Tricyclics Bearing Imidazoline, Oxadiazole, or Tetrazole Groups. *J. Med. Chem.* **1996**, *39*, 1857−1863. Smith, P. W.; Whittington, A. R. Novel inhibitors of influenza sialidases related to zanamivir. Heterocyclic replacements of the glycerol sidechain. *Bioorg. Med. Chem. Lett.* **1997**, *7*, 2239−2242.
(66) Fincham, C. I.; Higginbottom, M.; Hill, D. R.; Horwell, D. C.; O'Toole, J. C.; Ratcliffe, G. S.; Rees, D. C.; Roberts, E. Amide bond replacements incorporated into CCK-B selective "dipeptoids". *J. Med. Chem.* **1992**, *35*, 1472−1484.
(67) E.g., Jung, F.; Delvare, C.; Boucherot, D.; Hamon, A.; Ackerley, N.; Betts, M. J. Synthesis and structure−activity relationship of new cephalosporins with amino heterocycles at C-7. Dependence of the antibacterial spectrum and β-lactamase stability on the pK_a of the C-7 heterocycle. *J. Med. Chem.* **1991**, *34*, 1110−1116. Thompson, S. K.; Murthy, K. H. M.; Zhao, B.; Winborne, E.; Green, D. W.; Fisher, S. M.; DesJarlais, R. L.; Tomaszek, T. A., Jr.; Meek, T. D. Rational Design, Synthesis, and Crystallographic Analysis of a Hydroxyethylene-Based HIV-1 Protease Inhibitor Containing a Heterocyclic P1′−P2′ Amide Bond Isostere. *J. Med. Chem.* **1994**, *37*, 3100−3107.
(68) E.g., Pitt, B. Diversity of calcium antagonists. *Clin. Therap.* **1997**, *19(Suppl. A)*, 3−17. Ferrari, R. Major differences among the three classes of calcium antagonists. *Eur. Heart J.* **1997**, *18(Suppl. A)*, A56−A70. Triggle, D. J. Calcium-channel antagonists: mechanisms of action, vascular selectivities, and clinical relevance. *Clev. Clin. J. Med.* **1992**, *59*, 617−627. Triggle, D. J. Sites, mechanisms of action, and differentiation of calcium channel antagonists. *Am. J. Hypert.* **1991**, *4*, 422S-429S.
(69) Sadowski, J.; Wagener, M.; Gasteiger, J. Assessing similarity and diversity of combinatorial libraries by spatial autocorrelation functions and neural networks. *Angew. Chem., Int. Ed. Engl.* **1996**, *34*, 2674−2677.
(70) Schwarz, M. K.; Sauer, W. H. B. Manuscript in preparation.

CI025599W