Survey of the Diversity Space Coverage of Reported Combinatorial Libraries

Sara H. Fitzgerald, Michal Sabat, and H. Mario Geysen*

Department of Chemistry, University of Virginia, Charlottesville, Virginia 22904

Received December 1, 2006

Courtesy of the annual collections reported by Roland E. Dolle in the *Journal of Combinatorial Chemistry*, all three-point diverse libraries reported in the literature since 1992 have been evaluated according to their similarity at the library level (the Diversity Space approach).[1] This comparison enabled the identification of several particularly promising scaffold hopping opportunities and highlighted a number of optimal libraries (surrogates) expected to reveal binding information characteristic of an entire area of chemical space. As highlighted herein, future library design pursuits would benefit from a methodology such as the Diversity Space approach to ensure access to novel areas within the chemical landscape, thereby avoiding the expenditure of additional resources to cover a previously explored region.

Introduction

In their infancy, the focus of combinatorial chemistry methods was on quantity and speed of synthesis. Thus, libraries were chosen on the basis of established synthetic protocols and readily available monomer sets, with an emphasis solely on maximum diversity. From a drug discovery perspective, however, it was quickly realized that sheer numbers were not the solution to the problem because the diversity introduced in these initial libraries was often irrelevant or only partially relevant to the targets of interest and frequently disregarded the associated pharmacokinetic and pharmacodynamic properties.[2,3] It became clear, then, that a move toward more targeted or directed library design was warranted, but even now, chemists in the field continue to struggle with how synthetic priorities should be established. The pharmaceutical industry has provided a wealth of information regarding what it means for a molecule to be druglike. However, as highlighted by Ecker and Crooke, there is reason to doubt the value in the development of libraries based solely on these types of established motifs.[4] Rather, combinatorial chemists need a method by which to ensure that the compounds resulting from their syntheses are both pharmaceutically relevant *and* sufficiently novel. Thus, it is worthwhile to consider library design strategies that provide access to uncharted regions of chemical space, thereby replacing the previous emphasis on *diversity within libraries* with a newfound focus on *diversity between libraries*. This transition requires a substantially different approach than the molecular level comparisons offered by most of the currently available similarity measures.

As established previously, the Diversity Space methodology enables a quantitative assessment of similarity/diversity at the library level rather than the molecular level.[1] Specifically, comparisons are made with respect to the similarity or difference in the display of the diversity elements decorating each scaffold, resulting in a certain degree of coverage or overlap in diversity space. It is expected that information such as this may serve to improve chemists' library design decisions, particularly in terms of the practical applications that were reported in the previous publication, namely, scaffold hopping and surrogate synthesis.

To briefly review, the Diversity Space approach is applicable to scaffolds with three points of diversity, all of which are expected to contribute to the structure−activity relationship (SAR). The distances between a library's three diversity elements are combined to produce a *diversity triangle*, and the lengths of the sides of this triangle are further translated into the *x*, *y*, and *z* coordinates designating a point in *diversity space*. This space is divided into boxes to quantify the coverage of the space, and the degree of similarity or diversity between two libraries is assessed on the basis of the number of diversity space boxes the two libraries have in common (given as a percent overlap). As established, the Diversity Space methodology naturally lends itself to exercises that are central to the goals of combinatorial science. *Scaffold hopping* is a well-established approach, often employed to replace a component of a molecule that is undesirable (in terms of solubility, toxicity, binding affinity, etc.) or to avoid intellectual property (IP) infringements. To distinguish our approach to scaffold hopping from more traditional methodologies, we have classified it as "soft" scaffold hopping, highlighting the fact that a lack of molecular-level information may mean that alternative scaffolds identified in Diversity Space are not as absolute as those identified by other approaches. Despite this apparent disadvantage, however, the library context and the diversity included therein are expected to be more than enough to compensate for this "softness," although the results of future experimentation could conceivably invalidate this assumption. Additionally, to make better use of time and resources, *surrogate synthesis* designates the act of identifying the

* To whom correspondence should be addressed. Phone: (434)243-7741. Fax: (434)243-8923. E-mail: geysen@virginia.edu.

10.1021/cc0601579 CCC: $37.00
Published on Web 06/13/2007

optimal library to synthesize from a structurally related set of potential libraries. This surrogate library is chosen on the basis of the diversity space it shares in common with other members of the set, as well as on its own chemical tractability, which includes such significant factors as monomer availability, synthetic fidelity, and average yield.

The tenets and utility of the Diversity Space methodology were previously developed in the context of an in-house collection of structurally related libraries, but it was considered an appropriate extension to investigate the repertoire of combinatorial libraries in the public domain. In the case of the public domain collection, the synthetic approach and corresponding chemical tractability of each library has already been established. Given this synthetic advantage, it becomes even more beneficial to establish how these libraries compare with one another with respect to the coverage of chemical space. This type of assessment facilitates answers to the following questions: (1) Given a library or molecular structure of interest, to which libraries can I successfully expect to scaffold hop? (2) What are the best surrogate libraries to synthesize and screen to obtain information regarding the projected activity of several large subsets of libraries within the public domain collection? (3) How are combinatorial chemists performing in terms of exploring the universe of potential drug molecules? In other words, is this landscape being covered adequately, or are current libraries simply mimicking those of the past?

The survey of reported combinatorial libraries included herein has not been carried out in any type of systematic fashion, to date. The results not only suggest some obvious scaffold hopping and surrogate synthesis candidates but also address the degree of success with which combinatorial chemists have explored the chemical landscape over the past fifteen years.

Methods

The R−R distances for all three-point diverse libraries create a diversity triangle, the size and orientation of which represent the spatial display of the library's decorations. This diversity triangle can be mapped to a point in the three-dimensional environment of diversity space, a space which is subsequently partitioned into boxes to quantify the overlap of libraries. For the current study, then, the aim was to analyze all three-point diverse libraries that had been reported in the literature since the early 1990s, when combinatorial techniques first entered the chemical arena. Thanks to the diligent work of Roland E. Dolle, this task was simplified to a great extent, and all three-point diverse libraries were selected from his annual collections.[5−12] In some cases, four-point diverse systems were also chosen, but only three of the decorations were designated for the assignment of the library's diversity triangle. The selected structures, as well as any notes regarding decoration assignments or specific molecular modeling decisions that were made, are given in the Supporting Information. The name of each structure was assigned on the basis of the year in which the Dolle collection was published and the reference number from the Dolle collection that corresponds to the publication in which each library can be found. Thus, **99-10** indicates a library that was reported in the tenth reference of the 1999 Dolle paper, which surveyed the libraries from all of 1998. (Because there were two separate Dolle collections published in 1998, **98A** was used to designate the libraries from the collection with reported bioactivity[5] and **98B** was used to designate the libraries from the collection without reported bioactivity.)[6] A total of 698 libraries were chosen, and any undefined stereocenters were considered variable, resulting in a total of 1246 different scaffolds. The Markush structure of each scaffold was built in Catalyst 4.10, and to establish a reference for measurement of the diversity triangles, the decorations were modeled as methyl groups, with the carbon atom of the methyl substituent representing the point of attachment of the decoration to the scaffold. As established in the previous publication, this strategy reduces each library to the core structure common to all members, regardless of the identity of the decorations (assuming R_1, R_2, and $R_3 \neq$ H). It is important to note that the monomer sets for the public domain libraries were intentionally disregarded in this comparison because their further consideration would preclude the library level comparison desired. As such, however, any experimental validation or extension of the analyses presented should be preceded by a thorough consideration of the relevant synthetic parameters. As before, to account for the conformational flexibility of each library, conformational searches were completed in Catalyst 4.10 (BEST algorithm, maximum of 255 conformers, energy range of 10 kcal mol^{-1} from the global minimum conformation).

The sides of each diversity triangle are assigned to the *x*, *y*, and *z* coordinates of diversity space by starting at the 12:00 position and moving in a clockwise direction around the triangle. The rotational capability of each library thereby produces three nonredundant diversity space points for each conformation. It is important to note that this clockwise assignment is retained regardless of the numbers assigned to the R-group decorations in the library's Markush structure. The R-group numbering within the public domain set is understandably inconsistent because these numbers are often dictated by the chemistry on which each individual library is based. Maintaining a consistent clockwise frame of reference, however, ensures that the diversity triangles for the full set of public domain libraries can be appropriately compared.

As before, diversity space was divided into boxes (1 Å dimension on all sides) to quantify the coverage of the space, and the established overlap equations were used to create symmetric and asymmetric matrices depicting the library comparisons.

Symmetric

$$\% \text{ overlap} = 100[AB/(A + B - AB)] \quad (1)$$

Asymmetric

$$\% \text{ overlap}_A = 100(AB/A) \quad (2)$$

$$\% \text{ overlap}_B = 100(AB/B) \quad (3)$$

The % overlap values from the symmetric matrix can be used to scaffold hop between different libraries, a high value indicating similar coverage of diversity space and

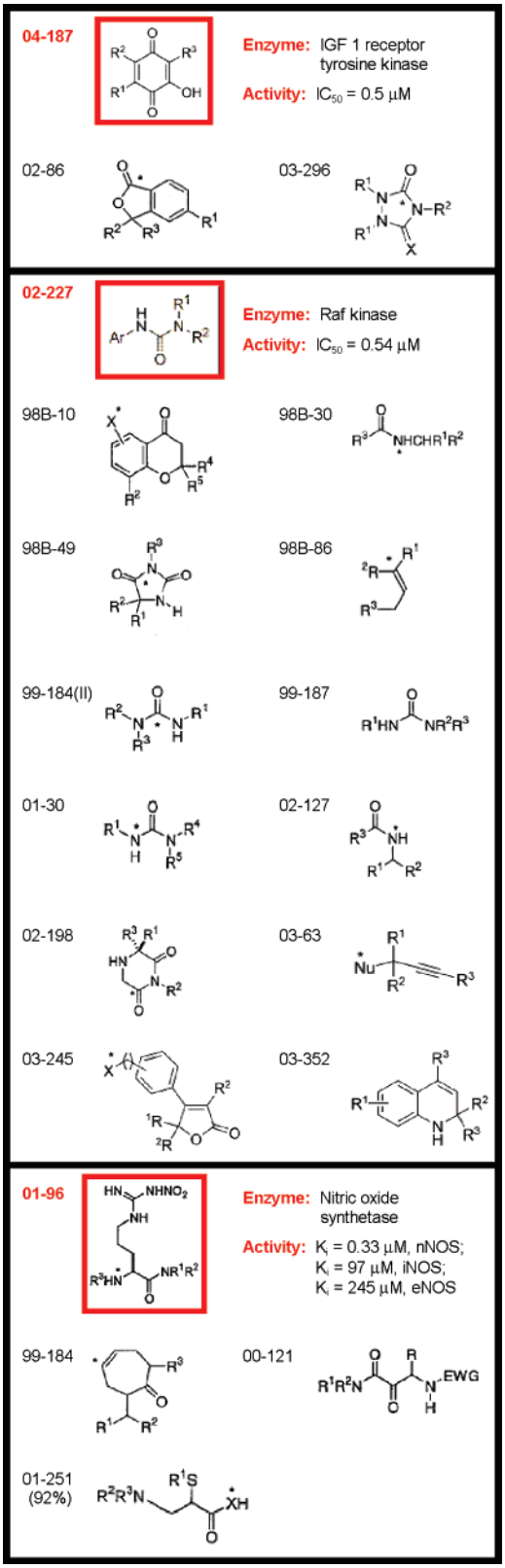

Figure 1. Illustration of the potential for scaffold hopping among several bioactive libraries in the public domain. In each black box, the bioactive library is highlighted in red along with the molecular target or enzyme against which it was evaluated and the potency of the most active library member (as given in the corresponding Dolle collection). The other libraries shown in each box display >90% overlap with the bioactive library in the symmetric matrix. If an exact percentage value is not given, it can be assumed to be 100%. (Again, any specific decisions regarding the modeling or decoration assignment for each of the depicted structures can be found in the Supporting Information.)

therefore an increased likelihood of success in scaffold hopping. The asymmetric matrix represents the overlap with respect to each library individually, thereby enabling the selection of an appropriate surrogate on the basis of the diversity space retained when one library is replaced with another. Thus, for the public domain collection, potential scaffold hopping and surrogate synthesis candidates were selected from the symmetric and asymmetric matrices, respectively.

As established previously, the Diversity Space methodology is most applicable when a small molecule can be reduced to a scaffold or core structure that is not expected to take part in the binding interaction with a protein target but is instead serving only as the *backside* template from which the interacting diversity elements or decorations protrude. Having not yet established a successful quantitative measure of this concept, however, we were not comfortable making a qualitative, and perhaps ambiguous, judgment as to which scaffolds fell into the "backside template" category and which did not. Thus, in the selection of three-point diverse systems from the public domain set, this front/back notion was not considered.

Results and Discussion

The full symmetric and asymmetric overlap matrices for the chosen public domain libraries are provided in the Supporting Information. (As established in the previous publication, the results for the stereoisomers of each library were collapsed, resulting in a 698 × 698 matrix for both the symmetric and asymmetric case.) Select examples from the two matrices are discussed in further detail below.

The symmetric matrix resulting from the survey of the full set of public domain libraries can be used to indicate potential candidates for scaffold hopping. In some instances, 100% overlap between two libraries indicated identical scaffolds that were either intentionally or unintentionally synthesized in multiple years. More interesting, though, are the instances of 100% overlap between scaffolds of nonidentical structure, indicating an exceptionally high degree of similarity in the spatial orientation of the decorations on the scaffolds despite the differences in their structures. This is of particular interest and importance when one considers the three-point diverse libraries for which accompanying screening data was reported (121 out of the 698 libraries surveyed). In this case, the value of scaffold hopping is evident; there is reason to believe that the biological activity witnessed in any one of the libraries may also be witnessed in each of its high overlap companions, provided that the same diversity elements could be employed (Figure 1).

For the asymmetric matrix, it was highlighted in the previous publication that surrogate synthesis is most applicable when one considers a structurally related set of libraries, all of which provide access to the same chemical space. Despite the fact that this type of "structurally related set" is not readily apparent within the set of public domain libraries surveyed here, it is still useful to consider how one might maximize the knowledge obtained from a single synthesis and screening. For example, 168 libraries were found to be complete subsets of library **00-137**, indicating a

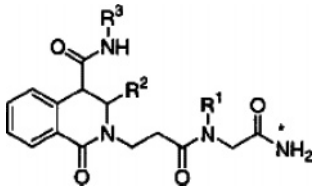

Figure 2. Markush representation of library **00-137**. Since there were two undefined stereocenters in this library, all possible stereochemistries were considered, and the conformers resulting from each stereoisomer were combined into a single file prior to analysis of the diversity triangles. Thus, **00-137** yielded 520 conformers (120 for **00-137RR**, 118 for **00-137RS**, 120 for **00-137SR**, and 162 for **00-137SS**). In the asymmetric matrix, 24.1% of the surveyed public domain libraries were found to be complete subsets of **00-137**.

profound promiscuity of this particular core structure (Figure 2). Of these 168 libraries, 24 were reported to have bioactivity against a variety of targets. Interestingly, no screening data was reported for **00-137**, although the value in assaying this particular library seems obvious from the results of this survey. Clearly, then, even though the public domain libraries are not structurally related enough to all provide access to the same area of chemical space, several appropriate surrogates can still be chosen, with the understanding that each one will only provide information about that portion of the full public domain set that is most similar to (or contained within) itself.

Further analysis of the data from the asymmetric matrix provides evidence for the performance of combinatorial library design to date. Averaging the values in each column of the matrix provides a measure of the average coverage of each library and enables a corresponding *surrogate rank* to be assigned. The coverage data for the top 50 libraries is shown in Table 1. When the cumulative number of complete subsets is graphed with respect to the surrogate rank, *half of the 698 libraries are already completely contained within the top 11 libraries* (Figure 3). Clearly, there exists a significant degree of redundancy in the diversity space that is being accessed by combinatorial libraries. Because of the widespread use of scaffold hopping and focused library synthesis, a slight degree of library-to-library similarity was expected in this survey, but not to as great an extent as it was found. To encourage better use of synthetic resources, then, future combinatorial pursuits need to break away from the spatial orientations contributing to this profound redundancy. Using the Diversity Space approach, chemists using combinatorial strategies have access to an effective means of comparing the coverage of libraries, thereby ensuring their proposed syntheses will tap into new areas of chemical space. To this end, it is important to know what proportion of this space is currently being covered. The dimensions of the diversity triangles for the public domain libraries ranged from 2.0 to 30.0 Å. Of course, after translating these dimensions to *x*, *y*, and *z* coordinates in diversity space and accounting for rotation, each axis of diversity space spanned an identical range. At the 1.0 Å box size, then, each axis was divided into 28 segments, resulting in 21,952 total boxes (28^3) to accommodate the diversity space points for the full set of public domain libraries. Of these, only 2688, or 12.2%, were actually filled, indicating that there are large regions of diversity space that remain to be explored. Of obvious concern is the upper limit, in this case, of 30.0 Å. Clearly, most small molecule libraries cannot be expected to achieve these distances nor are the dimensions necessarily applicable for the types of binding environments within which these small molecules are typically thought to interact. When the extent of coverage of the diversity space boxes of <10.0 Å

Table 1. Coverage Data for the Top 50 Surrogates from the Public Domain Set, Ranked in Order of Each Library's Average % Overlap in the Asymmetric Matrix of the 698 Libraries Surveyed

library	av % overlap	no. of complete subsets	cumulative no. of complete subsets[a]
00-137	44.96	168	168
99-51	39.64	124	194
00-230	38.90	137	215
99-4a	36.85	97	217
03-285	35.41	121	225
98B-15	34.94	116	237
99-26	34.59	94	302
98B-68oCO	34.53	104	303
03-131m	34.51	85	322
00-136	34.48	100	339
02-133	34.46	115	365
04-71	33.18	92	370
02-88	32.91	113	388
00-81	32.69	83	392
00-99	32.69	83	392
04-163	32.69	83	392
98A-27	32.66	82	396
98A-57	32.66	82	396
98B-248	32.66	82	396
03-282	32.66	82	396
01-50	32.12	96	397
03-131o	31.98	79	400
04-395	31.40	102	401
99-142o	30.58	75	406
99-142m	29.63	58	411
99-243	28.76	75	419
98A-9	28.61	40	420
02-73	28.34	67	422
01-243	28.31	101	459
98A-5	27.37	74	460
98B-113	26.91	73	464
02-3	26.82	64	465
02-5	26.72	66	469
02-177	26.43	56	470
03-221	26.27	73	471
98B-138	26.04	65	472
04-409	25.89	69	472
03-277	25.55	88	473
01-163	25.34	71	474
04-7	25.28	66	475
98A-28	24.80	75	476
99-22	24.43	92	486
04-324	24.00	63	487
02-194	23.87	55	489
02-7	23.85	61	490
02-126	23.25	58	491
04-385	22.92	59	495
00-251C	22.88	78	497
99-168	22.71	65	498
01-92	22.49	50	499

[a] The cumulative number of complete subsets represents the additional subsets obtained when the next ranked library is considered. In other words, of the 124 libraries that are complete subsets of library **99-51**, 26 of these are different than those covered by library **00-137**, bringing the cumulative number of subsets to 194.

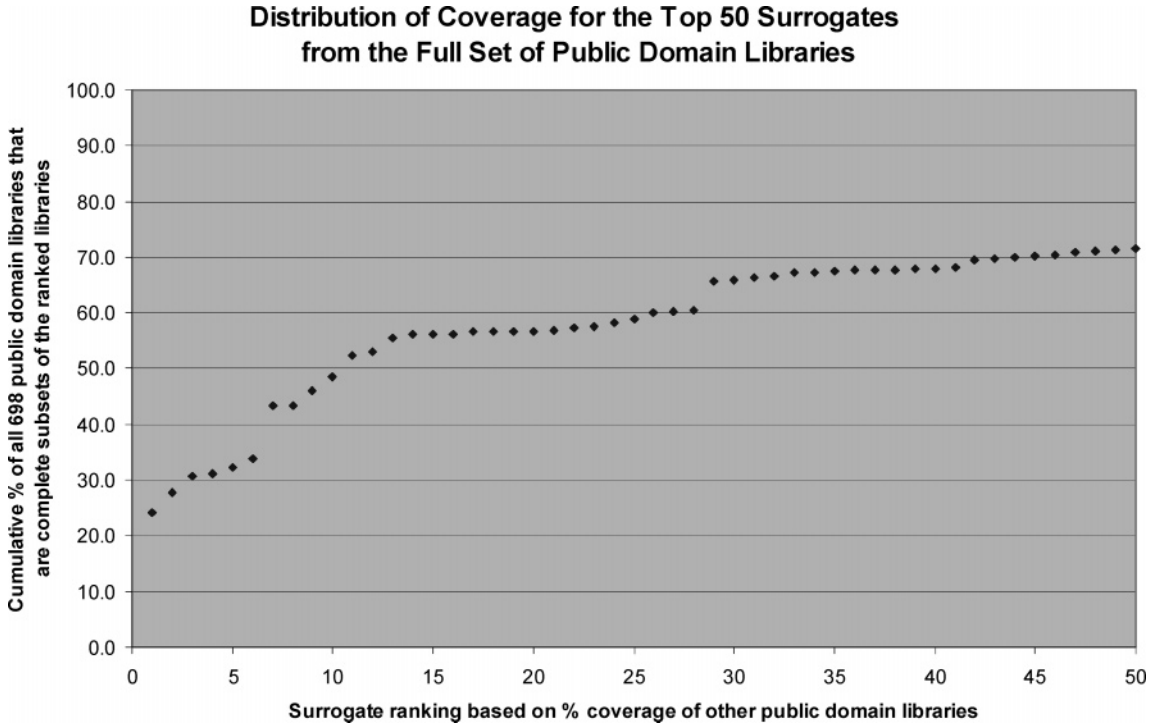

Figure 3. Graph showing the surrogate coverage of the top 50 Dolle libraries. The 1−50 rank was assigned on the basis of the average % overlap with respect to the other Dolle libraries. The cumulative number of libraries that were complete subsets of these top 50 was tabulated, and this was converted to a percent with respect to the total number of libraries surveyed (698).

was examined, the public domain set performed much more satisfactorily, filling 88.3% of the 512 total boxes (8^3). Likewise, for the diversity space boxes of <15.0 Å, 71.0% of the 2197 total boxes (13^3) were filled. As these dimensions are much more reasonable for small molecule library design, it is anticipated that Diversity Space may allow for the remaining 11.7 or 29%, respectively, of these regions to be identified and further explored.

It is important to re-emphasize here that, to achieve a library-level assessment, the Diversity Space methodology intentionally disregards the identity of the decorations on each scaffold in favor of a comparison based strictly on their spatial display. Thus, to say that two libraries exhibit 100% overlap with one another does not imply their absolute equivalency but rather their similarity with respect to the spatial display of their diversity elements. The resulting prediction of excellent scaffold hopping potential must therefore be followed by a consideration of the perceived synthetic approach and available monomer sets for each library. Likewise, in the selection of a surrogate library, it is not to be assumed that all of its subset libraries can simply be discarded from consideration. Rather, the selection of a surrogate or subset library should be made according to the extent of information known about the binding environment. The best surrogates tend to exhibit significant flexibility, enabling them to occupy large numbers of diversity space boxes. Because of this, however, the hits resulting from the screening of a surrogate library will likely be weak, as the enhanced flexibility will contribute to a significant entropic penalty upon binding. If very little is known about the binding environment, even these weak hits can reveal valuable structure−activity information, specifically with respect to the identity of the functionalities critical for binding. Once this preliminary SAR has been established, one of the surrogate's more rigid subset libraries can be employed, thereby avoiding the entropic penalty while allowing for the selection of the monomer set most appropriate to the established SAR.

A few other aspects of the Diversity Space methodology deserve some additional attention at this point.

Box Size and Boundary Location. Since diversity space is divided into boxes to quantify the coverage of the space, it is clear that the choice of box size and boundary location will change the number of diversity space points found in a given box. To establish the impact of these factors on the overall outcome, we decided to vary the box size and boundary location (currently implemented as 1.0 Å and integer boundaries) and reanalyze a small subset of the public domain libraries. The libraries designated as **98A** were chosen for this investigation, and the resulting symmetric and asymmetric matrices are shown in Figures 4 (varying box size) and 5 (varying boundary location).

As expected, when the box size is reduced, the corresponding degree of overlap is also reduced. In fact, by the time one reaches 0.25 Å, the only significant overlap that survives in either matrix is that of libraries **98A-27** and **98A-57**. Upon further inspection, it quickly becomes clear why this overlap is maintained because the two libraries turn out to be identical in structure. This example reveals how the diversity space box size can be fine-tuned according to the level of specificity desired. Depending on the investigation at hand, one may want to consider only libraries with a large degree of similarity, in which case a box size smaller than 1.0 Å would produce a matrix with very few high overlap pairs, thereby providing a more efficient selection filter. On the other hand, when it is worthwhile to know both the high

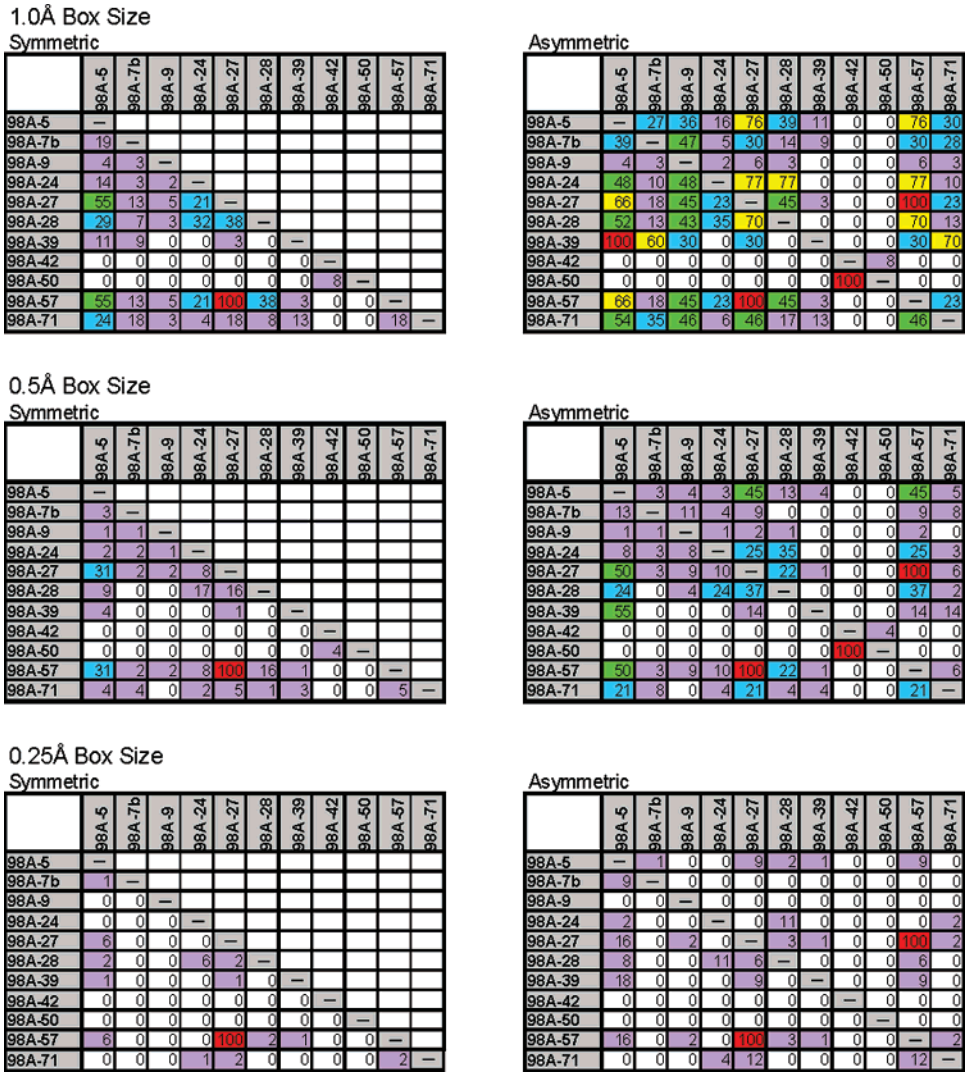

Figure 4. Symmetric and asymmetric matrices depicting the overlap of the **98A** libraries with a diversity space box size of 1.0, 0.5, and 0.25 Å.

and medium overlap libraries, a box size of 1.0 Å is sufficient. In terms of surrogate selection, in particular, it may be worthwhile to analyze the diversity space overlap with a larger box size. In this way, the average % overlap value that is used to rank the surrogates will not be swamped by libraries that may have identical counterparts in the set. For instance, at the 1.0 Å level, library **98A-5** has an average overlap of 42.9%, while libraries **98A-27** and **98A-57** have an average overlap of 43.5%. At the 0.25 Å level, library **98A-5** has dropped back to an average overlap of 6.9%, while libraries **98A-27** and **98A-57** have an average overlap of 13.6%. Although any of the three libraries appear to offer appropriate surrogate selections at the larger box size, the potential utility of **98A-5** disappears at the smaller box size, primarily because the averages for **98A-27** and **98A-57** are heavily biased by their 100% overlap with one another. Clearly, then, for cases where identical libraries may be present in the set under consideration, larger diversity space box sizes are more suitable for surrogate selection.

While varying the box size was expected to contribute to a corresponding change in overlap, varying the boundary location was anticipated to have only a marginal effect. Table 2 shows the surrogate ranking results for the **98A** libraries with different boundary locations. As before, this rank was assigned on the basis of each library's average % overlap in the Integer, Integer + 0.2 Å, and Integer + 0.5 Å asymmetric matrices. Although the mid-ranking libraries are shuffled a bit as the boundary location is altered, the top- and bottom-ranking surrogate libraries are clearly identified regardless of the boundary. As evident, then, from the surrogate rank as well as from the actual % overlap values given in the matrices of Figure 5, the slight overlap changes that result from the variation in boundary location do not in any way impair our ability to judge the quality of a library for scaffold hopping or surrogate synthesis. Since it is these more practical applications (based on relative comparisons and not the actual values) with which we are concerned, the seemingly arbitrary choice of boundary location is not expected to detract from the utility of the Diversity Space methodology.

Quantitative Approach to Front/Back Concept. As mentioned previously, when this survey of public domain libraries was begun, we did not have an adequate way of assessing the front/back nature of a library to determine whether the Diversity Space approach was relevant. Having established in the previous publication that the geometric

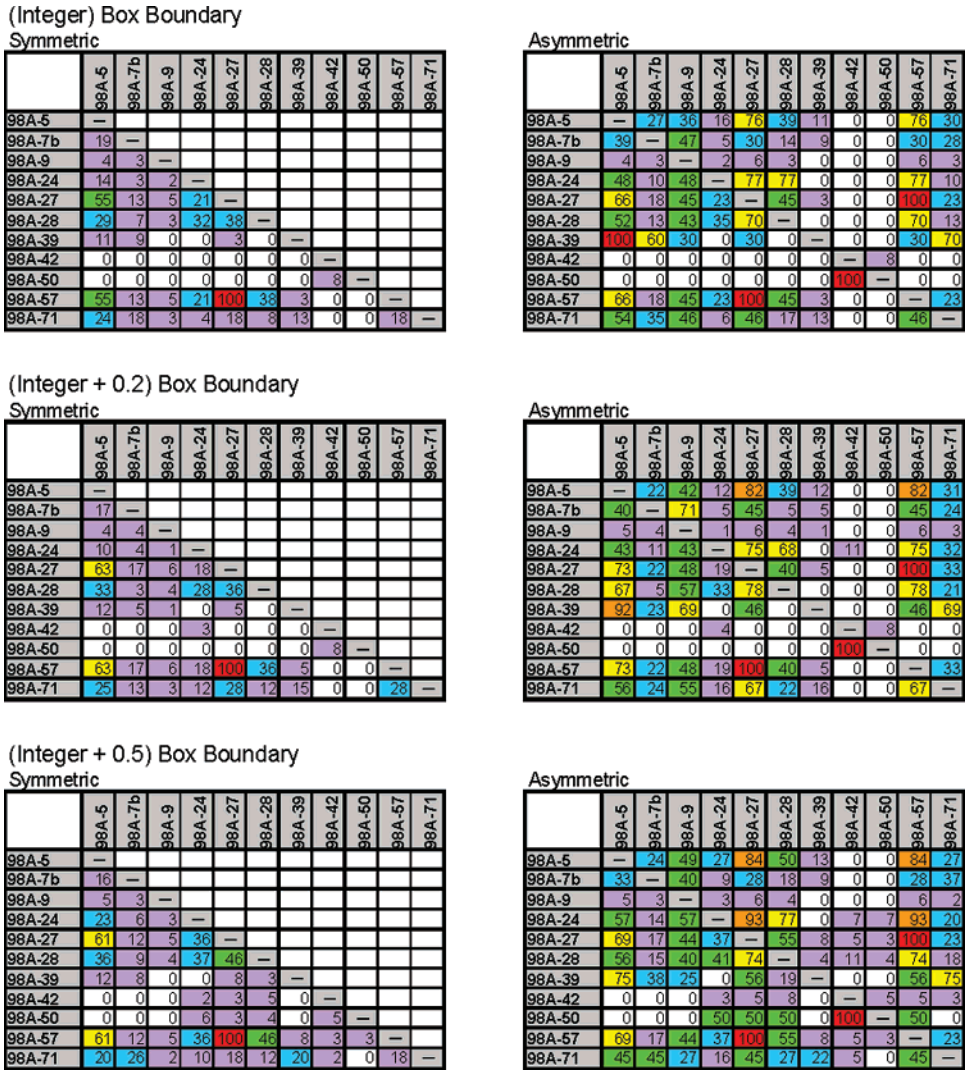

Figure 5. Symmetric and asymmetric matrices depicting the overlap of the **98A** libraries with a diversity space box boundary of Integer, Integer + 0.2 Å, and Integer + 0.5 Å.

Table 2. Surrogate Ranking Results for the **98A** Libraries with Varying Boundary Location[a]

Surrogate Rank	Integer Box Boundary	Integer + 0.2Å Box Boundary	Integer + 0.5Å Box Boundary
1=2	98A-27 98A-57	98A-27 98A-57	98A-27 98A-57
3	98A-5	98A-5	98A-5
4	98A-9	98A-9	98A-28
5	98A-28	98A-71	98A-9
6	98A-71	98A-28	98A-71
7	98A-7b	98A-7b	98A-24
8	98A-24	98A-42	98A-7b
9	98A-42	98A-24	98A-42
10	98A-39	98A-39	98A-39
11	98A-50	98A-50	98A-50

[a] The colors are used to distinguish more clearly between each library. Colors that are maintained across an entire row indicate rankings that are not affected by the choice of boundary location.

center of the library was not a good reference point from which to ascertain the vector directions, we turned instead to the diversity triangle itself. If the decorations on a library all contribute to the SAR and therefore constitute a contact surface, with the remainder of the scaffold serving only as a backside template, it was proposed that the plane formed by the diversity triangle would thereby define a "frontline" for the library, such that very little of the structure would extend beyond this plane. To establish the adequacy of this approach, the diversity triangle for every conformer of a library was placed in the *xy* plane. According to the "frontline" notion, most of the remainder of the structure should lie on the $+z$ or $-z$ side of this plane, such that the diversity elements are exposed and accessible from the opposite side. Thus, we analyzed the distribution of atoms in the *z* direction, and a conformer was considered "appropriate" for diversity space analysis if less than 25% of its atoms projected toward the front contact surface (on the opposite side of the *x,y* plane from the atom majority). Overall, a library was considered appropriate for Diversity Space analysis if at least half of its conformers were appropriate. An atomic distribution example is shown in Figure 6a for library **98A-7b**, a library for which the Diversity Space approach would seem highly applicable (27 of 29 conformers were appropriate). Alternatively, as shown

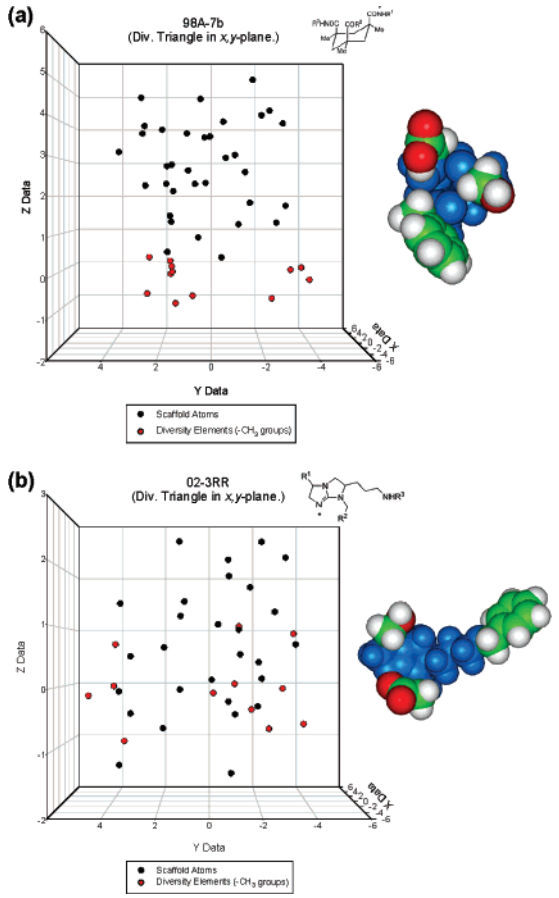

Figure 6. Atomic distribution for a single conformer of (a) library **98A-7b** and (b) library **02-3**, with the diversity triangle aligned in the *x*,*y* plane. The atoms making up the decorations (represented as methyl groups in the diversity space analysis) are shown in red in the atomic distributions to highlight the accessibility of these decorations. For a qualitative illustration, the two libraries are also depicted as space-filling models (scaffold shown in blue) with typical R-group functionality, represented by the side chains of serine, aspartic acid, and phenylalanine. The distribution of atoms reveals an obvious front contact surface in the case of **98A-7b** and no such accessible surface for **02-3**. As such, the Diversity Space approach is considered much more applicable for **98A-7b** because its scaffold appears to serve as the desired backside template rather than as a critical component to the binding interaction. (Note: The representative conformer shown in panel b is an R,R stereoisomer of library **02-3**, for which 40 out of 42 conformers were ruled inappropriate. However, because 155 out of the 169 total **02-3** conformers were inappropriate, it is clear that none of the other three stereoisomers resulted in an improvement in the front/back nature of this library.]

in Figure 6b, the atomic distribution for library **02-3** revealed it to be much less relevant for Diversity Space consideration (only 14 of 169 conformers were appropriate). For the sake of comparison with a qualitative assessment of front/back, Figure 6 also shows a space-filling model for the lowest-energy conformer of each of these libraries. It is clear from these models that the **02-3** scaffold is much more likely to be a component of the binding interaction rather than just a backside template from which the interacting diversity elements protrude. Clearly, though, a pictorial representation such as this is not practical for large numbers of libraries with multiple conformers, highlighting the value of a suitable quantitative assessment. Unfortunately, the atomic distribution approach retains a slight bias toward large structures because the 25% threshold is much less restrictive if more

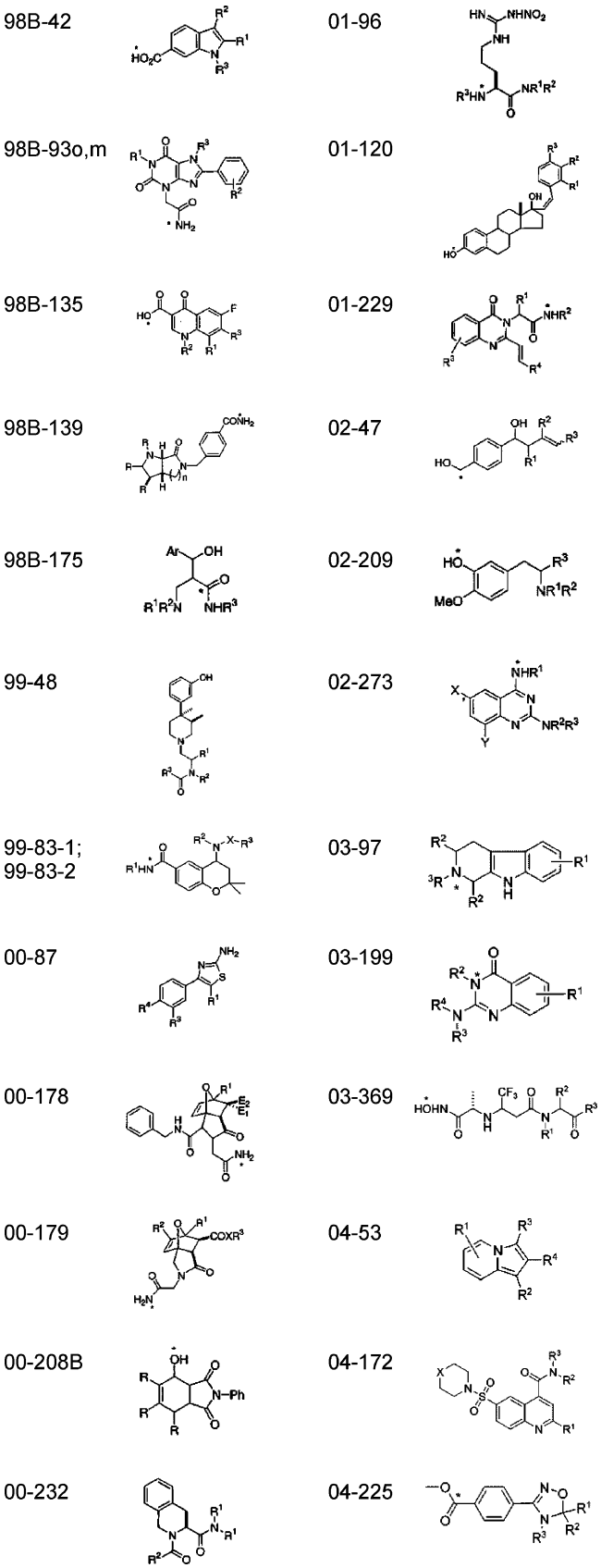

Figure 7. Select examples of public domain libraries with qualitatively *and* quantitatively acceptable front/back arrangements. (As before, any specific decisions regarding the modeling or decoration assignment for each of the depicted structures can be found in the Supporting Information.)

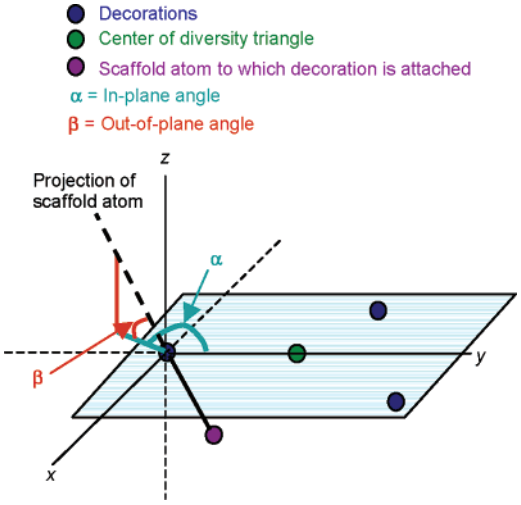

Figure 8. Illustration of the in-plane and out-of-plane scaffold atom (SA) projection angles. To calculate these angles for a given decoration, the diversity triangle is aligned in the *x*,*y* plane as shown, with the center of the triangle on the +*y* axis and the decoration of interest at the origin. The in-plane angle, α, determines the *x*,*y* quadrant of the scaffold atom projection, and the out-of-plane angle, *β*, determines whether the scaffold atom projects above or below the plane of the diversity triangle.

atoms are present. At this point, we have not developed an effective methodology that circumvents this bias. We continue to study this aspect of the Diversity Space approach, but for now, cautious use of the above strategy is suggested, keeping in mind the following caveats: Those libraries that are ruled to be appropriate for the Diversity Space approach *are* appropriate and can be analyzed without hesitation. Those libraries that are ruled to be inappropriate or less applicable for the Diversity Space approach should be examined further. A qualitative judgment may enable selection of several libraries that are indeed relevant but were too small to perform well according to the above criteria. Planar structures should only be considered appropriate for the Diversity Space analysis if all three diversity elements project from the same side of the 2D Markush structure. Otherwise, it is difficult to imagine that the scaffold would not somehow be involved in the binding interaction.

The quantitative front/back results for the 698 public domain libraries surveyed here are available upon request. A total of 147 libraries were found to successfully satisfy the above front/back criteria, some of which are shown in Figure 7. (For libraries with multiple stereoisomers, the library was ruled appropriate if any one of its stereoisomers was found to be appropriate.)

Spatial Orientation: Angular Component. As established in the previous publication, the tenets of Diversity Space are focused on the distances between a library's decorations rather than the angular component of their spatial orientation. However, in an attempt to find a better vector to define the front/back concept, several angles were found to be useful as a second tier of library comparison information, despite the fact that they did not significantly enhance the front/back analysis as initially anticipated. As shown in Figure 8, these angles define the projection of diversity from a given scaffold, and as such, we chose to call them the *in-plane* and *out-of-plane scaffold atom (SA) projection angles*.

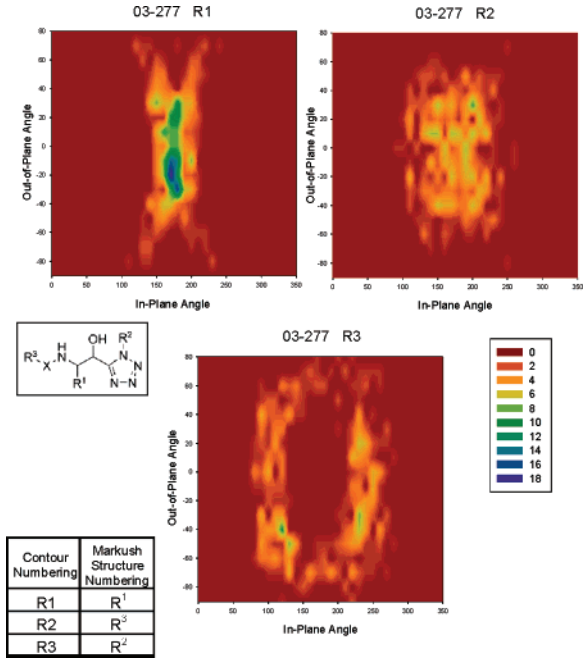

Figure 9. SA projection contour plots for library **03-277** (as produced in SigmaPlot 2001). As implied in Figure 8, the projections for each of a library's three decorations must be considered separately, thereby producing three different contours. Although the rotational capability of a library implies that the assignment of diversity space points can begin at any of the three decoration positions, the SA projection contour plots can accommodate only one absolute assignment of R1, R2, and R3. The Markush structure shown was taken directly from the corresponding Dolle collection. As discussed earlier, however, the Diversity Space methodology disregards the numbering of the original Markush structure so as to ensure a consistent clockwise frame of reference. Thus, the contour assignment of R1, R2, and R3, as it relates to the numbering of the Markush structure, is shown in the table in the bottom-left portion of the figure. (Note that although the same color scale is used for each of a library's three contours, this scale may or may not be the same as that of another library to which it is being compared.)

Because they define where a decoration is pointing in three-dimensional space, these angles are appropriate values to supplement the distance information previously obtained. Understandably, the projection from each decoration on a library is different, and unlike the distances, they cannot be collapsed into a compact representation such as a diversity space point. Nonetheless, it is worthwhile to consider this angular component, particularly to serve as a secondary filter for library selection. For instance, if several libraries present themselves as good scaffold hopping candidates on the basis of the symmetric overlap matrix, the angular information can be used to further optimize the selection, in the hopes of synthesizing the library that most closely matches the decoration spatial orientation, both the distance *and* angle, desired.

Given that the angles cannot be reduced to as compact a representation as the distances, how should we display this information? Clearly, the various conformers of a given library will produce multiple scaffold atom projections, and though it would be impractical to consider each of these independently, a contour plot enables a useful visualization of the projections for an entire library, where the in-plane and out-of-plane angles are designated on the *x* and *y* axes,

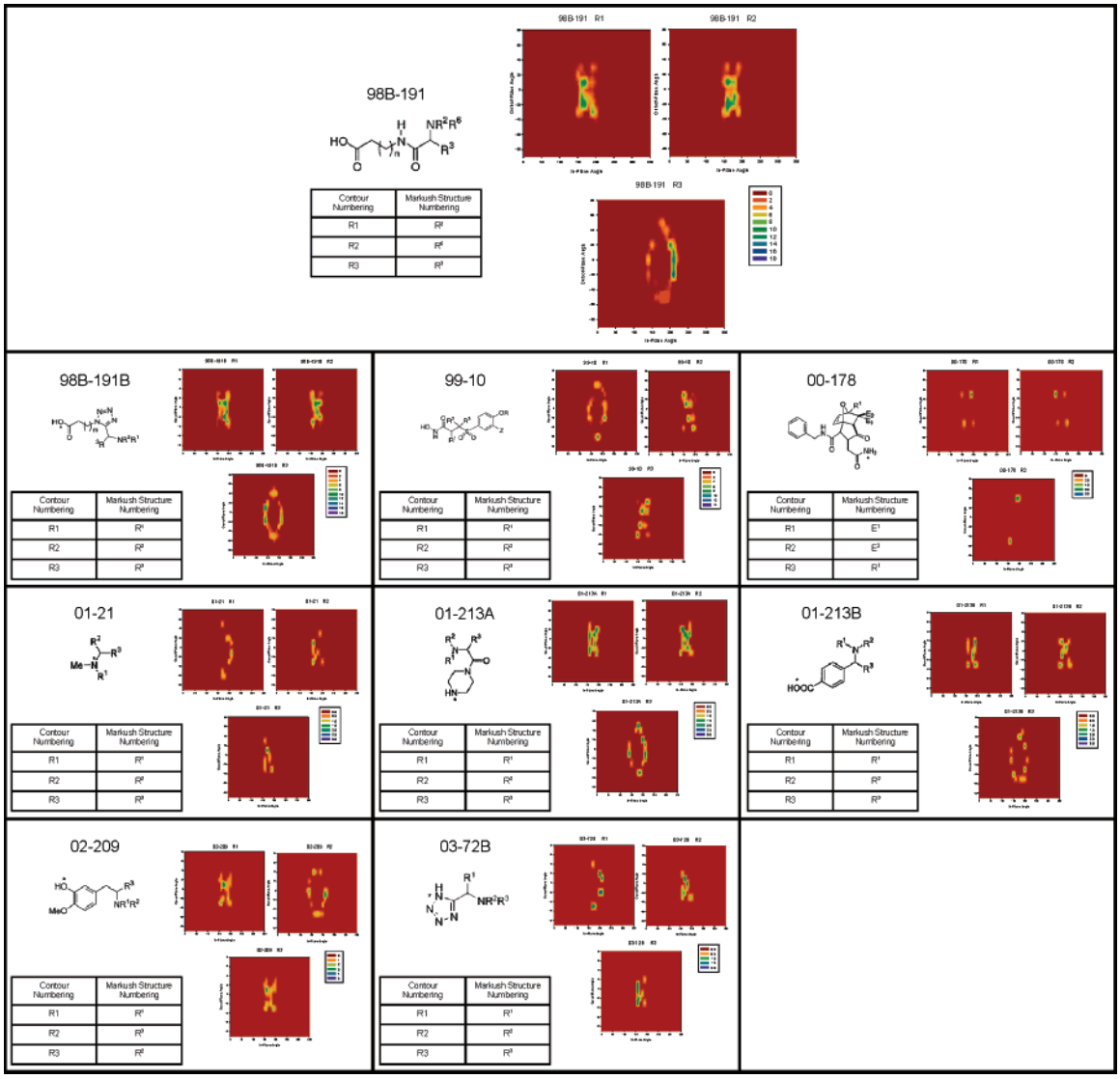

Figure 10. Illustration of the use of SA projection contour plots as a secondary filter for library selection. All libraries shown display a 100% overlap with library **98B-191** in the symmetric overlap matrix. Libraries **98B-191B**, **99-10**, **01-231A**, **01-213B**, and **02-209** have SA projections that also closely match that of library **98B-191**. Thus, they would be better choices for scaffold hopping than libraries **01-21** and **03-72B** (qualitatively intermediate angular overlap with **98B-191**) or library **00-178** (qualitatively low angular overlap with **98B-191**). As before, a table is included for each library to indicate the relationship between the Markush structure numbering and the assignment of R1, R2, and R3 for contour production. Because only a single assignment of R groups can be accommodated in the SA projection contour plots, there are cases in which the R1 contour in library A does not match up directly with the R1 contour in library B. However, when a library's rotational capability is taken into account, angular overlap can still be achieved as long as the R1/R2/R3 combination of contours in library A matches up with either the R1/R2/R3, R2/R3/R1, or R3/R1/R2 combination of contours in library B.

respectively, and the contour color represents the frequency with which this combination of angles occurs (Figure 9). As an added advantage, the use of this type of representation also provides a clear illustration of the flexibility afforded to a particular region of a library's structure. For example, if a large distribution of angles is seen for R1 and a small distribution for R2, one can assume that the R1 decoration sits on a more flexible region of the scaffold than does the R2 decoration. In most cases, the degree of flexibility witnessed in the contour plots appropriately complements our intuition regarding chemical structures.

To illustrate how one might use the SA projection contour plots as a secondary filter for library selection, consider the example shown in Figure 10. Library **98B-191** displays 100% overlap with eight other public domain libraries in the symmetric matrix. Considering only the distance component of the spatial orientation, then, any of these eight would appear to be good candidates for scaffold hopping. However, in looking at the SA projection contours, it is clear that some of these libraries match better than others with **98B-191** when the angular component is considered. Clearly, then, although the distance information should be considered the predominant overlap factor, the angular SA projection contours can be used as a qualitative secondary filter in cases where multiple libraries appear to be equally promising synthetic prospects. The SA projection contour plots for all 698 libraries surveyed have been produced using SigmaPlot 2001. However, because of the large number and hard-to-distribute format, these contours have not been included in the Supporting Information but are available upon request.

Conclusions

The Diversity Space analysis of the full set of three-point diverse public domain libraries revealed many interesting scaffold hopping and surrogate synthesis opportunities, a few of which have been highlighted here. More importantly, though, was the finding that, at the 1.0 Å box size, over half of the 698 libraries surveyed were already completely contained within the top 11 ranked libraries. Clearly, there is a need for an appropriate library design filter that will enable future synthetic pursuits to break away from the scaffolds and corresponding spatial orientations contributing to this redundancy. It is our hope that the Diversity Space methodology employed here will meet this need, enabling more informed library design decisions and enhancing the impact of combinatorial technologies on the drug discovery arena.

As was emphasized in the previous publication, Diversity Space is largely intended as a priority assessment tool, increasing the probability of synthesizing a library designed to meet a certain need, be it mimicking the coverage of another library or providing access, via a relatively unexplored spatial orientation, to a novel area of chemical space. Of course, because of its disconnect from such issues as monomer selection, yield, and synthetic tractability, the Diversity Space approach is not intended to be applied in isolation but should proceed instead with the input of trained combinatorial and medicinal chemists, whose insight may enable further prioritization of the results of a Diversity Space analysis. In addition, as with most computational methodologies, the general utility of the Diversity Space approach will be established by its use over time. With access to both a wide range of targets as well as extensive in-house library collections, research groups within the pharmaceutical industry are perhaps in the best position to establish the utility of Diversity Space in an objective and comprehensive manner. It is only after a thorough experimental survey such as this that we, or others, will be able to progress with the modifications and improvements necessary to make this a truly applicable tool. For now, we continue to explore the utility of this approach in a variety of practical settings, and we wait with anticipation for the dialogue that may be generated within the combinatorial field as the methodology is experimentally applied and thereby fine-tuned.

Supporting Information Available. The public domain libraries selected for Diversity Space analysis, as well as the resulting symmetric and asymmetric overlap matrices, are provided. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Fitzgerald, S. H.; Sabat, M.; Geysen, H. M. *J. Chem. Inf. Model.* **2006**, *46*, 1588−1597.
(2) Borman, S. *Chem. Eng. News* **1999**, *77*, 33−48.
(3) Oprea, T. I.; Gottfries, J.; Sherbukhin, V.; Svensson, P.; Kuhler, T. C. *J. Mol. Graph. Model.* **2000**, *18*, 512−524.
(4) Ecker, D. J.; Crooke, S. T. *Nat. Biotechnol.* **1995**, *13*, 351−360.
(5) Dolle, R. E. *Mol. Diversity* **1998**, *3*, 199−233.
(6) Dolle, R. E. *Mol. Diversity* **1998**, *4*, 233−256.
(7) Dolle, R. E.; Nelson, K. H., Jr. *J. Comb. Chem.* **1999**, *1*, 235−282.
(8) Dolle, R. E. *J. Comb. Chem.* **2000**, *2*, 383−433.
(9) Dolle, R. E. *J. Comb. Chem.* **2001**, *3*, 477−517.
(10) Dolle, R. E. *J. Comb. Chem.* **2002**, *4*, 369−418.
(11) Dolle, R. E. *J. Comb. Chem.* **2003**, *5*, 693−753.
(12) Dolle, R. E. *J. Comb. Chem.* **2004**, *6*, 623−679.

CC0601579