Chemistry Space Metrics in Diversity Analysis, Library Design, and Compound Selection

Paul R. Menard,* Jonathan S. Mason,† Isabelle Morize, and Susanne Bauerschmidt

Computer-Assisted Drug Design, Rhône-Poulenc Rorer, Collegeville, Pennsylvania 19426

Received June 6, 1998

DiverseSolutions software was used to generate a "universal" chemistry space that can be used as a standard for profiling most structural sets of interest. A nonlinear method for assigning structures to bins within chemistry space descriptors was developed. This allows the use of chemistry spaces scaled to include all structures within a set, while maintaining a reasonable distribution of structures within bins and providing target percentage cell occupancies. The universal chemistry space and nonlinear binning method were validated using random structures extracted from the Beilstein database. The approach was then used, in conjunction with other diversity analyses, for diverse subset selection and comparison of compound collections.

INTRODUCTION

Many, if not most, pharmaceutical companies use or are investigating the use of high-throughput screening (HTS) in their biological assay programs. This method is capable of processing hundreds or even thousands of compounds per day. To meet the increased demand, companies are testing their entire corporate compound repositories and are seeking to synthesize or acquire additional compounds. Combinatorial chemistry techniques[1,2] provide one method for synthesizing large numbers of compounds quickly. Whether obtained by synthesis or external acquisition, it is important to know if new compounds duplicate those already in-house, either in terms of structural characteristics or physicochemical properties. It may be desirable to duplicate the physicochemical properties of "drug-like" and/or different structural types of small molecules. Some screens require the use of lower numbers, leading to the need for subset selection. A number of methods for the diversity analysis of compound collections and related operations have been reported[3] to address these issues.

One such method is DiverseSolutions software,[4] developed by Professor R. Pearlman at the University of Texas at Austin. In this paper we describe the use of DiverseSolutions (DVS) to generate and identify combinations of parameters useful for analyzing and comparing a wide variety of chemical structures as well as the development of a nonlinear cell-based analysis method which allows the inclusion of compounds having outlying values for these parameters while maintaining a reasonable distribution of structures within the diversity space. The use of DVS and nonlinear binning is then illustrated with two applications: diverse subset selection and the comparison of compound collections.

Background. DiverseSolutions is a suite of tools used for calculating atomic/molecular descriptors, choosing a combination of these descriptors to provide optimum differentiation between compounds, and performing diversity-related tasks such as comparison of compound collections, nearest-neighbors analysis, and representative or diverse subset selection. Most operations can be performed either by using the graphical interface or by accessing the individual executable files from the command-line level. Two major versions of DVS have been released:[5] 2.x and 3.x. Version 3 offers several notable improvements over the previous release, including improved algorithms for descriptor calculation and chemistry space selection as well as better performance. The results presented here involve only the version 2.x code.

Many DVS functions require that a specific set of descriptors, known as a chemistry space (CSP), be defined prior to performing diversity operations. The chemistry space can be classified as either high- or low-dimensional. A high-dimensional CSP is typically defined by molecular fingerprints, e.g., bit strings describing the two-dimensional (2D) structural features present in a molecule, although any desired data can be included in the fingerprint.[6] Fingerprints must be calculated elsewhere and must be fixed-length; DVS provides the means for importing them, generating bit and nearest-neighbor statistics, and performing distance-based subset selection.

A low-dimensional CSP consists of a small number of molecular descriptors, typically 3 to 6, and is defined by three factors: type of descriptors, number of descriptors, and range of values for each descriptor. Descriptors may be derived from 2D structures (e.g., bonding patterns, atomic properties such as charge, or hydrogen-bond donor or acceptor ability) or 3D structures (geometric data such as interatomic distances, atomic properties, or whole-molecule data such as surface area). Precomputed descriptors containing any data considered relevant to diversity can be provided by the user. Alternatively descriptors can be calculated using DVS from either 2D or 3D structures; an integral version of CONCORD[7] is provided for 2D to 3D conversion. DVS descriptors include those based on BCUT values and those derived from SAVOL (molecular surface area and volume) and AM1 (MO energies, etc.) calculations; both these methods are also included in the DVS program suite.

† Current address: Computer-Assisted Drug Design, Bristol Myers Squibb, P. O. Box 4000, Princeton, NJ 08540.

10.1021/ci9801062 CCC: $15.00
Published on Web 10/14/1998

BCUT values[4] are an extension of work by Burden,[8] who sought to develop molecular identification numbers for organic structures, useful for such purposes as indexing structures in a database. The initial work dealt with constructing matrices containing the atomic numbers along the diagonal and numeric bonding data in the off-diagonal elements. Burden's molecular identification (MID) number was formed from the two lowest eigenvalues of this matrix. Researchers at Chemical Abstracts Service found that the MID showed signs of promise in that similar compounds were often assigned similar MID numbers, but, not surprisingly, they found that this one-dimensional chemistry-space metric was insufficient for general similarity searching purposes.[5]

As implemented in DVS, the diagonal elements may contain properties derived from 2D or 3D structures which are related to ligand−binding site interactions, such as atomic charge or H-bond donor or acceptor ability. On the off-diagonals are optionally data related to bonding (e.g., numeric bonding data), geometry (e.g., inverse interatomic distance), or other properties. The lowest and highest eigenvalues are considered, with the relative contributions from diagonal and off-diagonal elements modified by scaling factors, leading to over 60 BCUT descriptors per structure.

An autoselection feature is available for determining the best chemistry space for differentiating a given set of structures. The user specifies the descriptors to be considered, the maximum number to be selected, and a maximum percentage of input structures which may be excluded as outliers, which determines the range of values chosen for each descriptor. The range of each descriptor under consideration is subdivided into a number of equal bins, which are then populated. Correlated descriptors and those showing an uneven distribution of structures are eliminated, and a χ-squared approach is used to select the N descriptors which form the N-dimensional chemistry-space which best represents the diversity of the given set of structures.

A low-dimensional chemistry space can be used with distance and cell-based analysis methods; DVS provides both. In the cell-based approach, each of the N axes is subdivided into M bins of equal size, yielding M^N hypercubic cells. The N coordinates of each structure place it within one of these M^N cells. For some diversity-related tasks (e.g., subset selection), the software automatically selects the optimum value of M for that task. For other diversity-related tasks (e.g., identifying and filling diversity voids), the user controls the resolution with which he examines chemistry space by specifying his choice of M. A target of 12−15% cell occupancy is recommended, both to keep the number of empty cells at a reasonable level and to limit the number of compounds per occupied cell. The cell-based approach is attractive because (1) relationships between structures are clearly defined by cell membership, (2) missing diversity, i.e., empty cells, can easily be identified, and (3) very large datasets can be readily handled.

RESULTS AND DISCUSSION

"Universal" Chemistry Space Generation. A particular set of structures is perhaps best differentiated by a chemistry space calculated exclusively for that set. However, this has certain drawbacks. Approximately 20K structures are recommended as the minimum for automatic chemistry space selection in order to provide acceptable statistical significance and dimensionality. Many sets of interest are smaller than this. Furthermore, the comparison of compound collections is an ongoing activity. This includes the comparison of external collections and proposed combinatorial libraries against reference sets already in-house as well as the comparison of libraries against each other during the design phase in order to select the most diverse. It is clearly not efficient to develop new chemistry spaces each time, and the comparison of sets requires the use of a common chemistry space. The use of a standard chemistry space also provides historical data that can be used as permanent documentation for a collection and for future comparisons.

Table 1. Chemistry Space Descriptors for Various Sets

1. MDDR (63K structures)	4. CORP1 + MDDR (163K)
atomic charge	atomic charge, electronegativity
H-bond acceptor, donor ability	H-bond acceptor, donor ability
atomic polarizability (2)	atomic polarizability (2)
2. CORP1 (100K)	5. CORP1 + MDDR + CL1 (275K)
atomic charge	atomic charge, electronegativity
H-bond acceptor, donor ability	H-bond acceptor ability
atomic polarizability (2)	atomic polarizability (2)
	molecular surface area
3. CL1 (100K)	6. All sets (628K)
atomic charge (2)	atomic charge (2)
H-bond acceptor ability	H-bond acceptor, donor ability
atomic polarizability (2)	atomic polarizability (2)

A study was thus begun to identify a universal chemistry space that would be applicable to any structural set of interest. Six structural sets were used for this work: (1) three corporate collections of proprietary compounds: corporate sets 1 (CORP1, 100K structures), 2 (CORP2, 75K) and 3 (CORP3, 225K), (2) corporate combinatorial library subset 1 (CL1, 100K), (3) MDDR database[9] (MDDR, 63K), and (4) NCI database[10] (NCI, 83K). Structures in the corporate sets were nonoverlapping; these sets were differentiated by stock levels of the compounds (CORP1 & 2) and by type of studies for which the compounds were synthesized or acquired (CORP3). The combinatorial library set contained structures from a variety of chemistries done at RPR's global research centers, and many were of higher molecular weight and size than those in the other sets. The MDDR was of particular interest because its structures are primarily drugs or compounds with known biological activity. These six sets could also be combined to give a wider range of structural types and numbers (e.g., Table 1).

The structures for each set were obtained in SDF file format,[11] converted to SMILES codes,[12,13] and subjected to cleanup and filtering routines designed to provide a set of standardized SMILES codes for those compounds suitable for bioassay. Details are given in the Experimental Section. Structures were converted into 3D using the DVS-provided version of CONCORD, and standard 2D and 3D (but not 3D/MO) metrics were calculated using DVS. Chemistry spaces were calculated for each set and various combinations using the autoselection option with default program settings, including eliminating up to 10% of the input structures as outliers during chemistry space generation. The maximum number of descriptors to include in the final chemistry space was initially set to 6; this number was lowered in increments of 1 if suggested by status messages issued by DVS. Among the goals of this study was to identify trends in chemistry

space characteristics, such as type and number of descriptors selected, scaling factors chosen, and descriptor range variations. Six selected results are shown in Table 1.

In general six descriptors were selected for sets greater than 100K structures, five otherwise (the smallest set was 63K). Additional studies with the same sets showed that no more than six descriptors would be chosen regardless of how high the initial limit was set. In some cases two descriptors of the same general type, such as atomic charge, were chosen for one chemistry space. In detail these differed by scaling factor, high vs low eigenvalue, and/or 2D vs 3D data.

The following descriptors were frequently chosen: atomic polarizability (6/6 studies), atomic charge (6/6), and H-bond acceptor (6/6) and donor (4/6) abilities. These represent four of the seven standard 2D/3D BCUT descriptor types. Other descriptor characteristics, in particular scaling factors, were somewhat variable, but no attempt will be made to explain the details. Ranges were determined by the individual descriptor values of each structure in the set as well as the percentage of outliers, which was kept constant at up to 10% of input structures. They varied as expected for each descriptor and for the same descriptor in different sets.

There is of course no objective way to choose the perfect universal chemistry space. We sought to select a space that would be general (able to handle varied structural types, molecular sizes, and properties), scaled to accommodate very large sets, and contain commonly occurring descriptor types, which ideally would be related to ligand−receptor binding. The chemistry space based on 628K structures (Table 1, study no. 6) fulfilled these requirements and was chosen as one standard. This set contained compounds from a wide variety of sources: in-house compounds from combinatorial and traditional syntheses that involved a wide variety of chemistries spanning decades of research and nonproprietary "drug-like"structures (MDDR and NCI), many of which have known biological activity. The descriptors, in addition to their selection in most chemistry spaces reported here and in other trials, were also intuitively satisfying to the medicinal chemists.

Study no. 5 was the only one showing molecular surface area as a descriptor. This set contained 60% traditional drug-like molecules and 40% library compounds. The drug-like molecules were generally of moderate molecular weight and size, whereas a significant percentage of the library compounds were larger and heavier. Surface area may provide differentiation between these two compound categories. This chemistry space was also retained as one of the standards and may be useful in analyzing sets containing a wide range of molecular weights and sizes.

Nonlinear Binning. It is recommended that up to 10% outliers be allowed during chemistry space generation.[5,14] The reasons for this are 2-fold. First, each descriptor range will be concentrated on the majority of structures, which are the nonoutliers, leading to a reasonable distribution of structures (that is, a significant number of structures per bin with no bin holding the majority). Second, by definition outliers are structures having extreme values of computed properties and therefore may represent unusual molecules with little interest as drug candidates.

This approach posed two problems. The corporate structural sets were already highly filtered prior to diversity

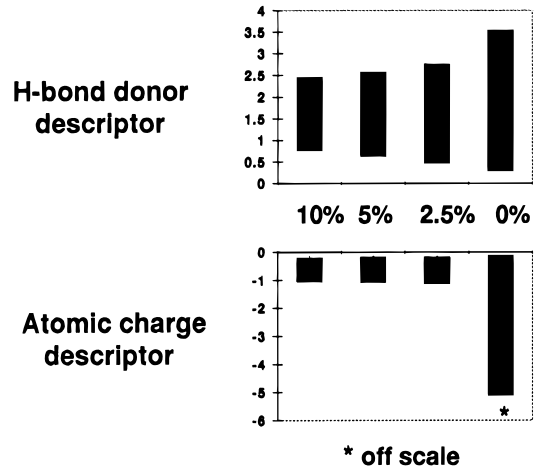

Figure 1. 628K chemistry space descriptor ranges vs percent outliers, 628K structural set.

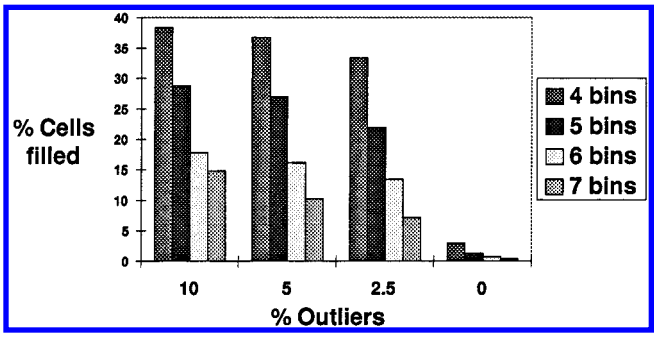

Figure 2. 628K chemistry space cell population vs percent outliers, 628K structural set.

analysis, and all remaining structures were considered interesting. If a nonzero percentage of outliers is chosen, those outliers will be definitively removed from further analysis or selection. Furthermore, visual inspection of outliers revealed them to be generally acceptable compounds for bioassay. In particular, all iodo compounds from the 628K set were found to be outliers in the polarizability metric using the 628K CSP calculated with 5% outliers. (This problem has been addressed in DVS version 3.x.)

A study was done to determine variation in descriptor ranges and percent cell population using the 628K structural set and chemistry spaces based on the 628K CSP but with the descriptor ranges extended to produce 5%, 2.5%, and 0% outliers rather than 10%. Results are shown in Figures 1 and 2. Four of six descriptors showed a significant but smooth increase in range from 10 to 0% outliers similar to that of H-bond donor ability (Figure 1) and were classified as narrow-range descriptors. The remaining two, both atomic charges, showed a similar trend from 10 to 2.5% outliers and then an extreme jump to 0% (wide-range). The more drastic example is shown in Figure 1. Cell occupancy was studied using the same outlier levels and from four to seven equal-size bins per descriptor (Figure 2). The recommended ideal occupancy range of 12−15%[5,14] could be reached by selecting the appropriate number of bins in all cases except 0% outliers. In the latter case virtually all structures were assigned to a single bin in each of the wide-range descriptors, resulting in extremely low overall cell occupancy.

DVS currently provides only a linear binning scheme for use with cell-based low-D chemistry space analyses: the software determines or the user specifies a number of bins,

and each descriptor in the chemistry space is subdivided into that many bins of equal range. The resulting cells are then populated, with the outliers being discarded. As one possible solution for retaining outliers, two additional bins could be added for each descriptor to hold structures below and above the minimum and maximum chemistry space ranges. However, this adds another $((M + 2)^N - M^N)$ cells, most of which will have little or no occupancy resulting in low overall percent occupancy. For example, for the 628K chemistry space using six bins per descriptor, the addition of two extra bins for outliers results in more than a 5-fold increase in cells.

Thus the nonlinear binning method was developed to allow the inclusion of outliers in a cell-based analysis, while maintaining an acceptable percentage of occupied cells and distribution of structures within bins. The general scheme is as follows:

(1) calculate descriptor values for a given set of structures;

(2) calculate or select a chemistry space for analysis scaled for x% outliers relative to this set of structures (x is typically 0−5%);

(3) calculate distribution histograms for each descriptor;

(4) create normalized coordinate files for all structures (i.e., files that contain the molecule IDs for all structures along with their values for each chemistry space descriptor scaled from 0 to 10);

(5) recalculate real space coordinate files from the normalized coordinates;

(6) starting from the distribution histogram as a reference, define a nonlinear binning scheme scaled to include all structures and provide a reasonable distribution of structures in the bins of each descriptor;

(7) perform binning and cell assignments for all structures, and calculate percent cell occupancy; repeat steps 6 and 7 if necessary until acceptable.

Steps 1−4 can be carried out from within DiverseSolutions. The nonlinear binning scheme can be designed using either normalized descriptor values, known as coordinates, as reported by DVS for a specific chemistry space (see Experimental Section), or raw data values obtained using an in-house conversion program. An advantage of the latter is that the binning scheme is independent of the chemistry space used to create the coordinate files, providing the same descriptors are used. Thus the same scheme can be used to analyze coordinate files originating from chemistry spaces having different ranges, and such files can be merged for a combined analysis.

The purpose of step 3 is to obtain reference distribution data that can be used as a template when designing the nonlinear binning scheme. This step assumes that the chemistry space being used is appropriate, i.e., that it was either defined specifically for this set of structures, has been shown to give reasonable distribution data, or is a universal or standard chemistry space. A nonzero percentage of outliers generally avoids compressing all structures into one or two bins in any of the descriptors. Numerous studies have shown that using 5% outliers with either a specific or the 628K CSP is a good starting point. Extra bins may be added in step 6 to maintain reasonable binning distributions while achieving desired cell occupancy.

Figure 3 illustrates the ranges and binning schemes for a narrow-range and a wide-range descriptor using the 628K chemistry space and either DVS linear (six bins, 5% outliers) or nonlinear (seven bins, 0% outliers) binning schemes. The nonlinear scheme was devised using the above approach, with an additional bin being added in step 6 and the bin ranges chosen so that a distribution pattern similar to the reference distribution was achieved. This is clearly evident in Figure 4, which shows corresponding changes in distribution of structures per bin for these schemes using the 628K set. Note that the nonlinear scheme (bottom graph) illustrates how reasonable distribution is maintained even with the inclusion of all outliers. Cell occupancy is 16.1% for the linear scheme (5% outliers) and 15.9% for the nonlinear, both of which are close to the target of 12−15%.

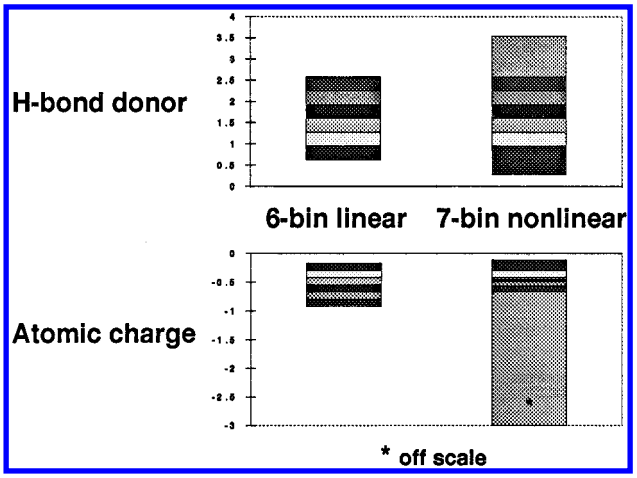

Figure 3. 628K chemistry space range and bins, linear vs nonlinear binning.

Universal Chemistry Space Validation. The purpose of this study was 2-fold: (1) to calculate chemistry spaces using structures unrelated to those described previously and compare the results to the universal chemistry space definition (628K CSP) and (2) to determine how well these new structures are handled by the universal chemistry space.

The Beilstein database[15] contains over 6 million organic structures which have been reported in the literature from 1779 to the present. These structures cover a broad range of chemistries and should well-represent accessible chemical diversity. RPR has an implementation of Beilstein in-house, and this database was chosen as a source of structures. Originally it was planned to identify, extract, and study all chemically reasonable organic structures. However, prototype studies showed that this would not be a trivial undertaking, performance considerations being a major factor. It was therefore decided to work with randomly selected structures taken from approximately 6.7 million organics that were identified. Five random sets of approximately 100K structures each were selected from all organics having a molecular weight of 1−1000 Daltons. These sets were also combined, and duplicate structures were removed, providing a single set of 450K structures. Chemistry spaces were calculated for the single and combined sets using conditions similar to those used in the universal CSP generation. Results are shown in Table 2.

Each of the 100K random sets gave chemistry spaces containing the same number (5) and types of descriptors. One of the five had a scaling factor variation for atomic charge and another one of five showed a similar variation for polarizability; other than that the descriptor definitions

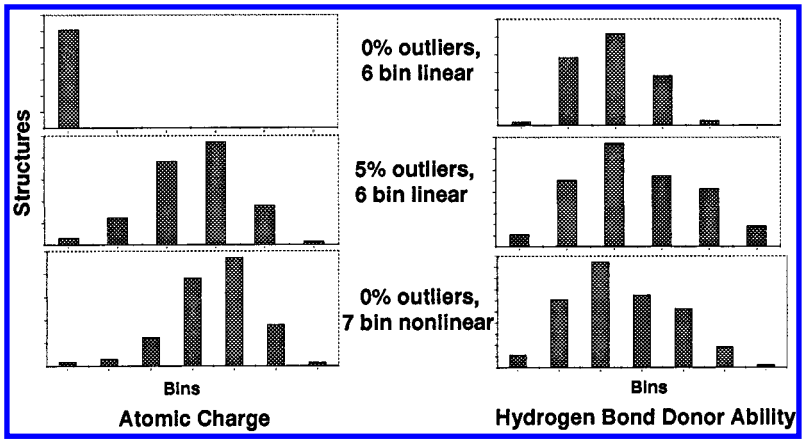

Figure 4. 628K chemistry space structure distribution (628K set), different binnings.

Table 2. Occurrence of Chemistry Space Descriptors in the 628K and Beilstein Random Sets

	set (descriptor occurrence)	
descriptor	628K	random_*x*
charge	2	1 × 5
H_acceptor	1	1 × 5
H_donor	1	1 × 5
polarizability	2	2 × 5

Table 3. Percent Cell Occupancies Using Various Binning Methods and Structural Sets

	set (% cell occupancy)		
binning method (% outliers)	628K	random_*x*	random_all
six-bin linear (5)	16.1	13.1	20.3
six-bin nonlinear (0)	28.1	22.1	34.0
seven-bin nonlinear (0)	15.9	12.6	20.9

were identical. For identical descriptors the ranges varied minimally (<5%). All five descriptor types are also present in the 628K set; minor scaling differences were noted in three, and a slight broadening in range was seen for one. Because of the good agreement between descriptors within the random sets and between the 628K CSP and them, it was concluded that the 628K chemistry space descriptors were applicable for general diversity studies on small organic molecules and consequently qualified as "universal".

The behavior of the random structures in the 628K CSP as a function of percent outliers was studied next. Each set was partitioned using the 628K chemistry space scaled for 5% outliers. An average of 11.4% of structures per 100K random set (11.3% for the combined set) fell outside this chemistry space range. This is clearly unacceptable for a chemistry space designed to give approximately 5% outliers regardless of structural set studied. Partitioning was then done using the 628K CSP scaled for 0% outliers. Less than 0.1% of structures were excluded for any set, which is acceptable for a chemistry space scaled for 0% outliers.

Finally, cell occupancy was studied using several binning methods. The separate and combined random sets, and the original 628K set, were partitioned into cells using (1) six linear bins and the 628K CSP scaled for 5% outliers, and (2) six and seven nonlinear bins with the 628K CSP scaled for 0% outliers. Results are presented in Table 3.

Although the six-bin linear scheme provides adequate cell occupancy, outliers are excluded. The six-bin nonlinear scheme gives excessively high percent cell occupation, while the seven-bin nonlinear scheme gives a reasonable cell occupancy and includes nearly all structures. It was concluded that the 628K CSP qualifies as a "universal" chemistry space and that nonlinear binning using the 628K CSP will be retained for future analyses. One should note that in all schemes from Table 3 the percentage of occupied cells is higher for the combined random set than for either the 100K individual random sets or the 628K set, even though the latter is larger. This indicates that (1) the same diversity is not represented by each of the 100K random sets (they are disjoint to some extent), and (2) there is greater diversity in the structures extracted from the Beilstein database than in our "drug-like" 628K set.

Representative Subset Selection. The previous studies have dealt with the identification of appropriate chemistry spaces and binning techniques which could be used for diversity operations on most if not all structural sets. In this section the use of these tools for subset selection will be described.

Many combinatorial libraries have been synthesized at RPR for testing against targets in the seven transmembrane G-protein coupled receptor (7TM-GPCR) class. These libraries were generally prepared in 96-well plates, with a single compound per well, and were then reformatted for biological testing and storage so that only 80 wells per plate were filled. Due to the large number of compounds it was desirable to select a smaller subset which could be given priority in new 7TM-GPCR-related bioassays. Since library compounds are generally tested on a plate-by-plate basis, a selection method would need to minimize the total number of plates chosen, rather than the total number of compounds. Furthermore, many of the libraries were designed to incorporate "privileged substructures", which are structural fragments chosen because of their frequent occurrence in known 7TM-GPCR-active compounds and their postulated importance to this type of biological activity. Thus the selection method should also be capable of choosing a diverse subset of compounds from each privileged substructure family.

Two approaches that were expected to be complementary were used. The first was based on use of the 628K CSP partitioning previously described. The second was based on four-point pharmacophores as implemented in ChemDiverse/Chem-X software[16] and customized and extended at RPR.[17] To present briefly the four-point pharmacophore method, a systematic analysis for all potential pharmacophores is performed for each structure with conformational sampling;

the pharmacophore is a combination of four center types and the six distances that position them in 3D space (a tetrahedron). Seven center types are available: H-bond donor, H-bond acceptor, aromatic ring, hydrophobe, acid, base, and one user-definable (default Chem-X = positively charged center, coded with the bases in our analyses). Center types within a molecule are assigned using atom types, including dummy atoms for aromatic rings and hydrophobic regions. Strategies for assigning atom types are described elsewhere.[17] For this study only six center types were used; the seventh was left undefined. Distances between center types are measured exactly in a molecule and are stored using a binning scheme, with each distance represented by the bin into whose range it falls. A nonlinear binning scheme was used, consisting of eight bins covering the range of 2−19.5 Å and two additional bins for distances falling below or above these limits.

All pharmacophores within a molecule are stored in a pharmacophore key, which is a bit string where each bit represents a specific combination of center types and distances. The key for a molecule is calculated by performing a conformational analysis and setting all bits corresponding to observed pharmacophores in all acceptable conformers. A customized conformational analysis method, using either random or systematic sampling based on molecular flexibility, that was developed at RPR was used. Logical operations such as AND or NOT can be performed using keys created under the same conditions and provide a means for calculating plate keys from individual molecule keys.

A three-step approach was used for subset selection. The first step used CSP cell-based partitioning with a selection scheme designed to sample by cell and privileged substructure while minimizing the number of plates chosen. The process is as follows:

−assign all compounds to cells using the desired chemistry space and binning method;

−partition each cell into subcells by privileged substructure, where each subcell contains only those compounds having the same privileged substructure (or lack thereof); multiple subcell membership is allowed;

−select the plate covering the most subcells;

−continue selecting plates covering the next largest amount of uncovered subcells;

−stop when all subcells are covered.

Output is a ranked list of plates along with the number of new subcells each plate covers.

The number of libraries considered was 31, and these were comprised of 67K compounds spread over 858 plates. Using the 628K CSP scaled for 0% outliers with the six-bin nonlinear binning method, a total of 649 out of 46 656 cells was filled, giving 1621 filled subcells. This corresponded to a low cell occupancy of 1.4%. This is not unexpected since we are focusing on libraries designed for a specific class of biological targets. As our goal was to reduce the number of plates needed for primary biological evaluation, selecting a higher cell occupancy would not have been a good idea since it would parallel with a larger number of plates to be selected. Three plate selections were made to cover 100, 95, and 90% of filled subcells, giving 218, 150, and 113 plates, respectively.

The second step involved augmenting these selections with plates chosen to provide four-point pharmacophore coverage.

Table 4. Plate Selection Using Four-Center Pharmacophores

% 628K CSP coverage	pharms present	pharms missing	% pharms missing	plates added[a]
100	1831K	366K	16.8	57
95	1690K	500K	23.0	54
90	1634K	553K	25.5	58

[a] To provide 75% coverage of missing pharmacophores.

Total pharmacophores represented by the library compounds were 2173K out of a possible 9700K. Pharmacophores covered by each of the three previous plate selections were calculated, and missing pharmacophores contained in unselected plates were identified (i.e., pharmacophores present in the library compounds as a whole but not shown by the compounds in any of the plates selected using the CSP-based approach). Depending upon the CSP-based selection, 74−83% of all pharmacophores present in the libraries were represented. A procedure was then developed to compare the pharmacophore key for each unselected plate (calculated by ORing the keys for all molecules on the plate) against the missing pharmacophores and rank them in order of coverage. A target coverage level of 75% of the missing pharmacophores was selected. This percentage would provide a good overall coverage of all library pharmacophores without causing substantially more plates to be selected. Results are shown in Table 4.

As noted in the chemistry space study, the 275K CSP was also retained as one standard for diversity assessment. It contained molecular surface area, a descriptor not present in the 628K CSP, and may differentiate structures in a manner different from the 628K CSP. Thus a check was performed to determine how representative our plate selection was in the 275K CSP, and a final step, similar to the first but using the 275K CSP scaled to include all structures, was carried out to select plates giving different percentage coverages of the filled cells in this chemistry space.

The atomic charge descriptor here was found not to differentiate these structures well (all structures were neutral and fell into the same bin) and was removed. To keep the number of additional plates selected to a relatively low number, a coarse, unsymmetrical nonlinear binning of either three or four bins per remaining descriptor was chosen. The entire set of plates was partitioned, and the ranking procedure was applied after considering the cells already covered by the plates selected in the first two steps. Coverage levels of 100, 95, and 90% were calculated. Final subset sizes ranged from 272 to 166 plates.

Comparison of Compound Collections. Another diversity-related application is the comparison of compound collections to provide, for instance, information on how similar or different the collections are. This information can then be used for either compound acquisition or in evaluating proposed libraries in order to select those that cover the most structural or property space that is not already represented by the corporate collection. Three approaches used routinely at RPR for diversity analysis and comparison of compound collections are based on 2D structural characterization (Daylight[18] fingerprints), atomic/molecular properties (DiverseSolutions) and 3D pharmacophores (ChemDiverse). In this section the use of all three methods to compare compound collections is discussed.

The collections examined were the MDDR database, the corporate collection CORP1, and the combinatorial library subset CL1, which represent a total of 275K structures. The goals of this study were many: to determine the overlap between collections using different analyses, to define diversity which is accessible by drug-like (MDDR) structures but not represented by any RPR compounds, to determine differences between the library and traditional compounds, etc. The results can be used in assessing commercially available compound collections for potential acquisition, in future library design, for selecting lead follow-up compounds, and in general for gaining insight into the strengths, weaknesses, and similarities between different approaches to diversity. Only enough results will be included here to illustrate the methodology.

Daylight fingerprints are generated from 2D molecular structures and are bit strings representing structural characteristics derived from an analysis of all bond paths containing typically between two to eight atoms. Nearest neighbors for each molecule are identified based on similarities of the fingerprints determined using the Tanimoto metric. The clustering method provided and used here is the Jarvis−Patrick nonhierarchical method, which states that two molecules will cluster together if (a) they are in each other's nearest neighbors list and (b) they have **J** of their **K** nearest neighbors in common. Clustering was done using nearest neighbors lists of 16, and values for **J** and **K** of 8 and 16. These values for the **J** and **K** parameters have been previously found to give reasonable average and maximum cluster sizes for various structural sets. This procedure is known to produce a significant number of singletons, many of which could logically be grouped into clusters of related structures.[19] There are various alternatives to dealing with the singletons. In this case the single rescue option was set to 4, which causes singletons to be assigned to the cluster occupied by the majority of their nearest neighbors, assuming that at least four belong to the same cluster. This produced 16K clusters and 2050 singletons (0.75% of total structures) from the 275K structures, as compared to the same number of clusters and 22 536 singletons (8.2%) without singleton rescue.

The DiverseSolutions analysis was done using the 275K chemistry space scaled for 5% outliers, which was derived from this total set of structures using the Auto-choose option. Six linear bins per descriptor were used, giving a cell occupancy of 14.8% (6914 occupied/46 656 defined cells).

The four-point pharmacophore analysis was done using six defined center types and the 10-bin distance classification described earlier. The total set contained 4300K pharmacophores. Individually the MDDR set was most diverse with 3800K pharmacophores, followed by CL1 (3200K) and CORP1 (3000K).

It is difficult to compare the effectiveness of one diversity method against another in differentiating between structures. The ability is dependent not only on the theoretical basis but also on the conditions used. For instance, DiverseSolutions results will differ for the same compound collection based on chemistry space, binning method, and number of bins per descriptor. Furthermore, since some diversity methods assign structures to logical groups such as clusters or cells, it is necessary to examine results based on both group and structure statistics. To illustrate the differences between the three methods used here, the number of entities common to the three sets as well as the ones specific to MDDR were determined and studied. Results are illustrated in Figures 5−7.

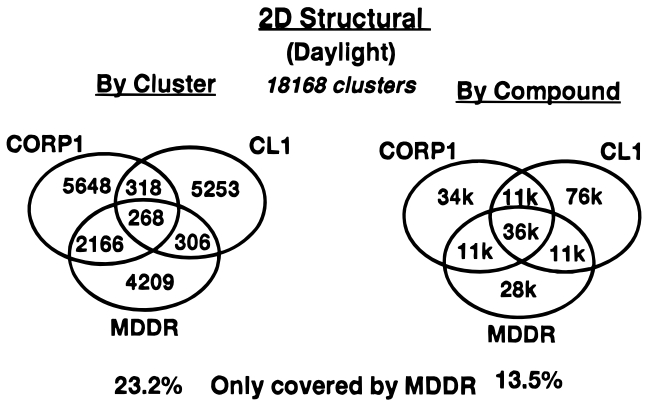

Figure 5. Structural comparison of compound collections.

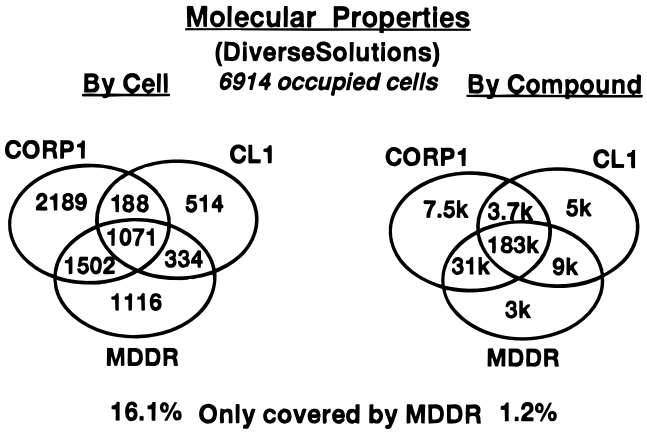

Figure 6. Property comparison of compound collections.

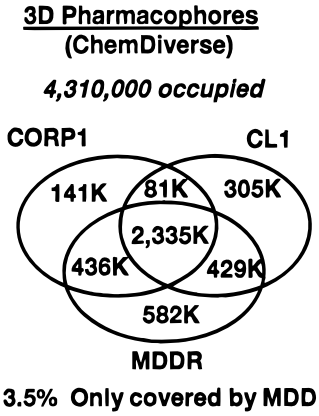

Figure 7. Pharmacophore comparison of compound collections.

The clustering results show that 1.5% of clusters contain structures from all three sets, comprising 17.4% of all structures. The large structure-to-cluster ratio indicates that these clusters represent common chemical families, which appear in all sets and are well-represented by individual structures. Clusters containing only MDDR compounds were 23.2% of the total, holding 13.5% of all structures. Here the structure-to-cluster ratio is smaller, showing that the MDDR contains many structural families not represented by the other sets, although the number of structures per family is rather small. For the DiverseSolutions approach 15.5%

of occupied cells contained representatives from all three sets, totaling 75.6% of all structures. This suggests that, although the total occupied cells span a wide range of chemistry space, most of the structures have relatively similar physicochemical properties regardless of origin. There were 16.1% of cells (1.24% of structures) specific to MDDR. This agrees somewhat with the clustering results: the MDDR contains structures having particular combinations of physicochemical parameters not seen elsewhere, but the number of representatives per family is small. Finally, 54.2% of total pharmacophores occur in all three sets; 13.5% are specific to MDDR. This is not surprising since a single molecule may contain several center types and is capable of expressing many pharmacophores. Thus all three methods provide different perspectives on how these sets differ and can be considered complementary; each can give useful information depending on the particular problem at hand.

CONCLUSION

DiverseSolutions software provides a powerful set of tools for diversity analysis using descriptors which combine 2D and 3D structural information with computed physicochemical parameters. The use of the nonlinear binning method as developed at RPR provides added flexibility when using cell-based analyses in conjunction with a low-dimensional chemistry space (the concept of nonlinear binning has now been implemented in the newest release of DiverseSolutions software, version 3.1.0). The combination of linear or nonlinear binning with a specific or universal chemistry space provides a general method for performing diversity operations on most structural sets of interest. These approaches, in conjunction with other diversity methods, have been effective in performing diverse subset selection and have provided insight into the similarities and differences between three compound collections that may be used for compound acquisition or the design of new libraries.

EXPERIMENTAL SECTION

Universal Chemistry Space Selection. The structures for all sets except the NCI collection were maintained as MACCSII 2.2 databases[9] and were extracted as SDF files. The NCI structures were obtained from the National Cancer Institute already in SDF format. SDF files were converted into SMILES codes using the program **mol2smi**, distributed by Daylight Chemical Information Systems.[18] Salts were neutralized, functional groups that could be represented in more than one form were standardized, and duplicate structures were removed using a series of in-house programs. Compounds unsuitable for bioassay (e.g., reactives and toxics) were removed as described in an earlier paper.[20] For the MDDR, dipeptide-like structures and structures having a molecular weight greater than 1000 Daltons were also removed.

DiverseSolutions software version 2.x was used for all studies. Structures were input as their processed SMILES codes; conversion to 3D was done using the on-board version of CONCORD. CONCORD settings, controlled by the initialization file.concordrc on a Silicon Graphics computer system running IRIX 5.3, were set to obtain maximum conversion levels [example settings: max_rotor 200, max_ring 50, max_at 500, ccr 0.25]. No attempt was made to further optimize 3D structures. Standard 2D and 3D, but not 3D/MO, metrics were calculated for all sets and various combinations of these sets. Default scaling factors (0.25−4.0) and H-suppressed matrices were used. Chemistry spaces were calculated using the Auto-choose option and up to 10% outliers. The maximum number of descriptors to consider was initially set to 6 and was reduced in increments of 1 when the number of compounds were determined to be too small as indicated by notification messages from DVS.

Nonlinear Binning. Chemistry space ranges were varied while keeping the descriptors fixed by selecting the "user selection of metric types" option, selecting metrics from an already defined chemistry space and associated set of molecules, and recalculating the chemistry space for different percentages of outliers allowed.

Normalized coordinates were obtained for a set of structures by choosing a chemistry space composed of the desired metrics and extending the ranges if necessary as described above to encompass all compounds. The option "report chemistry space coordinates for sample compounds" was selected, with the input list being the list of all structures. This produced a file containing one record per line, with each record containing a molecule ID and its values for each metric in the chemistry space, normalized from 0 (corresponding to the minimum value of the range for each descriptor) to 10 (maximum value). A program was developed to read this file and the corresponding chemistry space definition and produce a similarly formatted file containing raw data.

Nonlinear binning was performed using in-house code. This program took as input one file containing molecule IDs and either normalized or raw coordinates and a second file containing the number of bins and their ranges (per descriptor). Output was a file containing molecule IDs, bin assignments (an integer from 1 to M corresponding to the molecule's bin assignment for each descriptor), and cell IDs, made by concatenating the bin assignments. Histograms showing the distribution of structures vs values for each descriptor could be obtained using either the DVS option "generate histograms of raw metric values" or in-house code. Percentage of occupied cells was calculated for linear binning using the DVS option "generate cell statistics" or for nonlinear binning by using in-house procedures.

Universal Chemistry Space Validation. Version 96.2 of the Beilstein database was used. To identify organic structures the database was searched using Crossfire and the following combined queries, specified using the Substructure and Fact editors: (1) at least one carbon atom per structure (Draw "C", turn Free Sites on), (2) no isotopes allowed ("Exclude Isotopes" on, atoms D and T specifically excluded), and (3) pharmaceutically acceptable atoms only allowed (create an atom list of all elements but Li, B, C, N, O, F, Na, Mg, Si, S, Cl, P, Ca, Br, I, then "Exclude Selected Atom List"). For performance reasons searches were done by molecular weight range in increments of 25−500, spanning the values 1−2500. Large hitlists were further subdivided into files of approximately 100K records. Over 6.7 million database entries were identified using this procedure.

Five random sets of 100K database entries each were created by selecting from the hitlists corresponding to molecular weights of 1−1000. Random numbers were

generated and converted into hitlist names and entries within the hitlist, and molecule IDs were extracted. Structures corresponding to these IDs were extracted from the database in SDF format in increments of 2000, using utilities supplied by Beilstein. SMILES codes were created using **mol2smi**. The structures retrieved contained salt components and some unwanted atom types, so additional processing was required. Records having unwanted atoms were removed, salt components were removed, counterions were neutralized, functional groups were standardized, and duplicate structures were removed. Average set size after filtering was 94 558. All five sets were combined, and duplicates (approximately 5% of total structures) were removed, yielding 449 384 structures. Chemistry spaces were calculated for the individual and combined sets as described earlier using a maximum of 5% outliers.

Representative Subset Selection. The combinatorial libraries were prepared using 96-well plates, with a single compound per well, and were reformatted for biological testing and storage using 80 wells per plate. SMILES codes were generated for each structure using in-house routines, with the molecule ID consisting of the library number and a unique integer value corresponding to each compound in that library. Separate files were available to map each molecule ID back to a plate ID in cases where the molecules were not consecutively numbered. ISIS structural databases[9] were available for all library compounds. ISIS substructure searches were used to create lists of molecule IDs containing privileged substructures (one list per substructure). A separate list was maintained of those molecules containing no privileged substructures. Files containing molecule IDs and chemistry space cell assignments were generated as described under "Nonlinear Binning".

To perform the plate ranking using the cell-based approach with any chemistry space, the program "cherry" was developed. Input consists of files containing (1) molecule IDs with cell assignments, (2) privileged substructure names and either the library numbers or molecule IDs which contain them, (3) plate map information (molecule IDs mapped back to their plate IDs), and (4) optionally, a list of preselected plates. The program first makes the subcell assignments by appending a number corresponding to a specific privileged substructure (or lack thereof) to the cell number for each molecule. Molecules containing more than one privileged substructure are assigned to multiple subcells. The plates are then ranked in decreasing order by the number of uncovered subcells that they represent (preselected plates, if any, are ranked in order before all others). That is, the plate which contains the most filled subcells is selected first. A counter of selected (covered) subcells is updated, the plate containing the largest number of remaining uncovered subcells is selected, and these two steps are repeated until no uncovered subcells remain. Output is a file containing one record per line, with each record containing a plate ID, total number of subcells on that plate, number of covered subcells on that plate, and details of subcell coverage. Any desired percentage of subcell coverage can be obtained simply by selecting the appropriate plates from this ranked list.

Pharmacophore analysis and selection was done using Chem-X version JAN97. Databases of 3D structures were prepared from SMILES codes using a standalone version of CONCORD. Four-point pharmacophores were calculated using six of the seven possible center types (H-bond donor, H-bond acceptor, aromatic ring, hydrophobe, acid, base); the seventh was left undefined. Distances between centers were assigned using eight nonlinear bins spanning a total range of 2−19.5 Å; two separate bins were used for distances falling below or above this range, respectively. Center type definitions were assigned to atoms using a customized parameter file. Pharmacophore keys for each compound were calculated using either systematic or random conformational sampling, based on flexibility of the molecule, and were stored on a plate-by-plate basis (cumulative key for all structures on that plate). Keys corresponding to uncovered pharmacophores were created using logical operations (AND, OR, NOT) on the appropriate plate keys. Plates were ranked according to the number of pharmacophores their keys covered relative to a key of uncovered pharmacophores using scripts written in Chem-X Programmable Control Language (PCL).

Comparison of Compound Collections. Preparation of the MDDR, CORP1, and CL1 structures were described under "Universal Chemistry Space Selection".

Daylight software version 4.41 was used. Fingerprints were generated from SMILES codes using the **fingerprint** module, and nearest neighbors were calculated using **near-neighbors** with an upper limit of 16 neighbors. Clustering was done using **jarpat** with the clustering parameters set to 8/16 and the singleton rescue option set to 4. Cluster membership data was obtained using files produced by **showclusters**.

DiverseSolutions partitioning was done using the 275K CSP scaled for 5% outliers and standard linear binning using six bins per descriptor. Cell membership data was obtained using the DVS routine **cellnum**. Relationships between compound collections for the clustering and DVS analyses were determined by analyzing the cluster and cell membership data, respectively, using programs developed in-house.

Four-point pharmacophore analysis was done using six defined center types of seven possible and 10 distance bins, as described under "Representative Subset Selection." Relationships between compound collections were generally determined by performing logical operations on appropriate pharmacophore keys.

Programs developed in-house were used to compare the three sets and to produce the results shown in the figures.

REFERENCES AND NOTES

(1) Thompson, L. A.; Ellman, J. A. Synthesis and Applications of Small Molecule Libraries. *Chem. Rev.* **1996**, *96*, 555−600.

(2) Hermkens, P. H. H.; Ottenheijm, H. C. J.; Rees, D. Solid-phase Organic Reactions: A Review of the Recent Literature. *Tetrahedron* **1996**, *52*, 4527−4554.

(3) For example: (a) Warr, W. A. Combinatorial Chemistry and Molecular Diversity. An Overview. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 134−140. (b) Ashton, M. J.; Jaye, M. C.; Mason, J. S. New Perspectives in Lead Generation II: Evaluating Molecular Diversity. *Drug Discovery Today* **1996**, *1*, 71−78. (c) Willett, P. Using Computational Tools to Analyze Molecular Diversity. In *Combinatorial Chemistry*; *A Short Course*; DeWitt, S. H., Czarnik, A. W., Eds.; American Chemical Society: Washington, DC, 1997.

(4) Pearlman, R. S.; Smith, K. M. Novel Software Tools for Chemical Diversity. *Perspectives Drug Discovery Design* **1998**, *9*, 339−353.

(5) *DiverseSolutions User's Manual version 2.0.1* and *version 3.0.1*; Laboratory for Molecular Graphics and Theoretical Modeling, University of Texas: Austin, TX, copyrighted 1996, 1997; distributed by Tripos Inc., 1699 S. Hanley Road, St. Louis, MO 63144, 1996, 1997.

(6) Bradshaw, J. Adding Nonstructural Data into Fingerprints. Paper given at the 1998 Daylight User Group Meeting; Santa Fe, NM, February 1998.
(7) CONCORD software; Tripos Inc.: 1699 S. Hanley Road, St. Louis, MO 63144.
(8) Burden, F. R. Molecular Identification Number for Substructure Searches. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225−227.
(9) MDL Information Systems Inc.: 14600 Catalina Street, San Leandro, CA 94577.
(10) NCI Structural Database; National Cancer Institute: Rockville, MD.
(11) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244−255.
(12) Weininger, D. SMILES 1. Introduction and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.
(13) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.
(14) Pearlman, R. S. University of Texas at Austin, personal communication, 1997.
(15) Beilstein Information Systems, Inc. 14600 Catalina Street, San Leandro, CA 94577.
(16) Chemical Design Limited: Roundway House, Cromwell Park, Chipping Norton, Oxfordshire OX7 5SR, UK.
(17) (a) Mason, J. S. Experiences with Searching for Molecular Similarity in Conformationally Flexible 3D Databases. In *Molecular Similarity in Drug Design*; Dean, P. M., Ed.; Blackie Academic and Professional: Glasgow, 1995; Chapter 6, pp 138−162. (b) Mason, J. S.; Pickett, S. D. Partition-Based Selection. In *Perspectives in Drug Discovery and Design (PD3) − Special Issue on Computational Methods for the Analysis of Molecular Diversity*; Willett, P., Ed.: Kluwer Academic: 1997; Vols. 7/8, pp 85−114. (c) Pickett, S. D.; Mason, J. S.; McLay, I. M. Diversity Profiling and Design Using 3D Pharmacophores: Pharmacophore-Derived Queries (PDQ). *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 1214−1223.
(18) Daylight Chemical Information Systems, Inc.: 27401 Los Altos, Suite 370, Mission Viejo, CA 92691.
(19) Menard, P. R.; Lewis, R. A.; Mason, J. S. Rational Screening Set Design and Compound Selection: Cascaded Clustering, *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 497−505.
(20) Lewis, R. A.; Mason, J. S.; McLay, I. M. Similarity Measures for Rational Set Selection and Analysis of Combinatorial Libraries: The Diverse Property-Derived (DPD) Approach. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 599−614.

CI9801062