Diversity Profiling and Design Using 3D Pharmacophores: Pharmacophore-Derived Queries (PDQ)

Stephen D. Pickett,*,† Jonathan S. Mason,*,‡ and Iain M. McLay†

Computer-Aided Drug Design, Discovery Research, Rhône-Poulenc Rorer, Dagenham, Essex RM10 7XS, UK, and 500 Arcola Road, Collegeville, Pennsylvania 19426

Received June 19, 1996⊗

The current interest in combinatorial chemistry for lead generation has necessitated the development of methods for design and evaluation of the diversity of the resultant compound libraries. Such methods also have application in selecting diverse sets of compounds for general screening from corporate databases and in the analysis of large sets of structures to identify common patterns. In this paper we describe a novel methodology for calculating diversity and identifying common features based on the three-point pharmacophores expressed by a compound.[1] The method has been implemented within the environment of the Chem-X molecular modeling package (ChemDBS-3D), using a systematic analysis of 3D distance space with three point combinations of six pharmacophoric groups. The strategy used to define the pharmacophores is discussed, including an in-house developed atom type parameterization. The method is compared with the related approach being developed into the ChemDiverse module of Chem-X. Results from an analysis of a large corporate database and examples of combinatorial library profiling with both methods are presented. The use of 3D pharmacophores for assessing diversity, and the application of such methods to combinatorial library design, is discussed.

INTRODUCTION

In recent years, the pharmaceutical industry has become increasingly concerned with the area of molecular diversity, such as methods for diversity analysis within databases of chemical structures. The discovery of an innovative new chemical entity requires three key steps: the discovery of relevant biological targets, the generation of "lead" compounds, and the optimization of these leads. Lead generation and lead optimization are areas which may be addressed with suitable measures of molecular diversity and similarity. For lead generation such methods are of use in increasing the diversity of available compounds for biological screening through the synthesis of combinatorial libraries or the identification of externally available compounds. An appropriate measure of diversity allows the selection of diverse, representative, or focused sets of compounds from a database to generate screening sets. Similarly, it would be possible to identify compounds/libraries which complement, rather than duplicate, the appropriate properties already represented in the repository. For lead optimization the methods can be used to identify the key structural/property patterns from a large set of active/inactive compounds and to design, around these patterns, small targeted libraries to explore for new receptor/enzyme interactions.

Many molecular property descriptors are potentially useful for assessing the diversity or for clustering compound databases; some of these are listed in Table 1. The similarity metrics commonly used have tended to rely upon 1D and 2D properties of the molecules and/or fragment incidence data, for example molecular weight, *c* log *P*, topological indices, substructure "fingerprints" etc. Combinations of these properties have been used previously.[2−7] However, where 3D structures have been used, for example in the calculation of quantum mechanical properties[2] it was normally practical to deal with only single conformers. 3D distances (as used for example in 3D database feature−feature distance keys) can be used as a 3D metric for diversity, but they appear to show little improvement over 2D parameters, giving only a weak representation of the additional information on properties and shape available from 3D analyses.

Table 1. Some Descriptors Potentially of Use in the Assessment of Diversity within a Compound Library

structural information required	descriptor type
1D	molecular weight
2D	topological descriptors
	Daylight fingerprints, MACCS keys (structural characterizations)
	flexibility, shape (e.g., κ-indices)
	molecular property/physicochemical descriptors
	hydrophobicity (*c* log *P*)
	functional group or group property (e.g., H-bonding) counts
3D	quantum mechanical descriptors
	shape indices
	feature-feature distance keys (e.g., Chem-X, Unity)
	pharmacophores (≥3 centers)—e.g., pharmacophore-derived queries (PDQ), ChemDiverse

Enzymes/receptors recognize the shape and electronic properties of molecules and diversity methods need to take account of such properties. Given that a pharmacophore is a necessary condition for binding to a particular biological receptor, it can be considered to represent key aspects of bioactive shape and electronic properties. 3D pharmacophores with three or more centers and other shape descriptors begin to exploit true 3D information, but conformational flexibility needs to be taken into account to fully explore the potential 3D interactions of a molecule; analysis of single

† Rhone-Poulenc Rorer, Dagenham.
‡ Rhone-Poulenc Rorer, Collegeville.
⊗ Abstract published in *Advance ACS Abstracts,* October 15, 1996.

S0095-2338(96)00039-X CCC: $12.00

conformers can provide valuable 3D information, but conformational flexibility may be needed to ensure an increased yield over 2D diversity measures. The centers used in the pharmacophores need to be defined using property-related environments (e.g., hydrogen-bond donors and acceptors, acids, bases, hydrophobic/lipophilic and aromatic centers) rather than being atom-based. It is this desire to include both shape and property into the diversity calculation, allowing for conformational flexibility, which has led us to develop the method of pharmacophore-derived queries (PDQ), described in this paper.

When the challenge is the design of diverse or biased (e.g., to a particular receptor type) combinatorial library, aside from the choice of descriptor, the decision must also be made as to whether to profile the building blocks or the final structures. Analyzing building blocks can provide some data on diversity within a library, but it is very difficult to compare libraries; it is also difficult to identify missing diversity. Analyzing building blocks makes certain assumptions about the additivity of the descriptors once the groups are joined to the scaffold. This is not the case with a 3D descriptor applied to complete molecules, particularly once conformational flexibility is taken into account. The PDQ method described below is intended to be able to profile in 3D the final structures. This also makes the technique suitable for analyzing compound collections which have not been constructed in a combinatorial sense, for example corporate databases or collections of compounds available for purchase. The method may of course be used on building blocks, or derivatives thereof, as a preliminary design filter. The results of the PDQ profiling include structures superimposed upon the pharmacophore queries, enabling further computational studies to be performed.

The pharmacophore-derived query or PDQ approach is a novel method for diversity analysis, utilizing the three center pharmacophores present within a molecule as the descriptor. The method is based on searching 3D structural databases using a systematic coverage of pharmacophore types and sizes. The database can then be partitioned according to the pharmacophores that are covered and for each compound information on which pharmacophores are matched can be obtained; for each pharmacophore the number of times it is matched is also known. Central to its effectiveness is the perception of atom center types with associated properties, for example hydrogen-bond donors and acceptors, charged centers, acids, bases, aromatic rings, and hydrophobes. An approach similar in philosophy, though different in implementation, is currently being developed by Chemical Design Ltd. (ChemDiverse module of Chem-X[8]). A comparison of the two methods is described later.

1. DESCRIPTION OF PDQ STRATEGY

1.1. Overview. The PDQ strategy is based upon partitioning the compounds in a database of 3D structures in terms of the three center pharmacophores which are expressed by the compounds. An outline of the procedure is shown in Figure 1. The centers used in the pharmacophores were chosen to express groups commonly used in 3D searching and which represent important drug−receptor interactions: hydrogen-bond donor, hydrogen-bond acceptor, acid, base, aromatic center, and hydrophobe; distance ranges covering most expected pharmacophore sizes were used (2−24 Å).

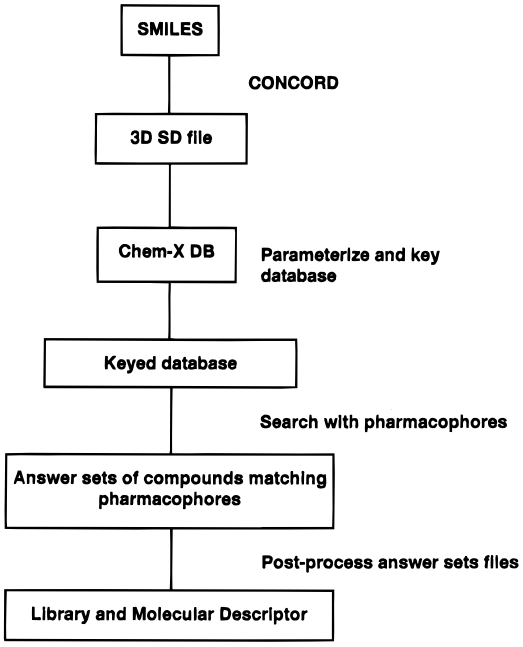


Figure 1. Flow chart of pharmacophore-derived queries (PDQ) methodology.

A systematic set of queries covering all possible pharmacophores (combinations of types and distances) is used to search the database. The search engine is ChemDBS-3D which allows extensive customization of the atom parameter file and several conformational search methods. In-house customization has permitted us to easily distinguish the different pharmacophore center types, ignoring "deactivated" atoms (e.g., substituted nitrogens in amides and aromatic rings) and identifying the many different acidic and basic environments, keeping them separate from other hydrogen bond donors and acceptors. For each unique combination of three pharmacophore groups, all permissible queries are generated covering combinations of six distance ranges: 2−4.5, 4.5−7, 7−10, 10−14, 14−19, and 19−24 Å. A total of 5916 geometrically valid queries are generated and stored within a Chem-X database. A series of control scripts have been written to generate and store the queries and to run the searches. Results are obtained as a series of "answer sets"—lists of compounds which hit each query. These lists are then processed to generate several results files and the molecular descriptor, a bit string indicating the queries hit by a compound. For each pharmacophore, it is possible to identify not only if a compound or database of compounds can express it but also how many times.

A difficulty with exploring the 3D properties of a molecule is that the properties change as the molecule explores available conformational space. The conformational flexibility of molecules has been taken into account through the use of a 3D searching system incorporating conformational flexibility (e.g., ChemDBS-3D). It has been shown to be important that the molecule not only satisfies the pharmacophore center-center distances but is required to fit the query within a defined tolerance, otherwise the shape of the hit conformation may not resemble accurately that of the pharmacophore.[9] ChemDBS-3D applies a fitting constraint subsequent to identification of a conformation satisfying the distance constraints.

1.2. Database Building. The compounds to be profiled are built into a ChemDBS format 3D structural database as

Table 2. Some Possible Atom Types Needed To Describe Different Nitrogen Environments with Associated Center Types[a]

14C: quarternary	3	14E: amide NH	1	14H: aromatic	2
14D: Nsp3 basic	3/7	14Q: amide NR2	-	14Y: tautomeric	1,2
14W: Nsp3 tert./quart.	3/7	14R: sulfonamide NH	1,2	14Z: arom. substituted	-
14U: Nsp2 basic (amidine)	3/7	14S: sulfonamide NR2	2	14F: pyrrole	1
14N: hydrazine NH	1	14B: imine	2	14A: nitrile	2
14P: hydrazine NR2	2	14M imine NH	1,2	14J: nitro	-
14V: aniline NH	1			14T: tetrazole	2/6
14X: aniline NR2	2				

[a] 1 donor, 2 acceptor, 3 positive charge, 4 aromatic, 5 hydrophobic, 6 and 7 only used with ChemDiverse for acidic and basic centers.

follows. SMILES[10,11] codes are generated for all the compounds in the library. Procedures have been developed in-house to automatically generate the codes based on R group definitions.[12] The SMILES are converted to a 3D SD file[13] using the program CONCORD.[14] The SD file is then read into Chem-X. Parameterization is performed at this stage to identify the different atom environments. Chem-X offers several options for keying the database according to the 3D feature−feature distances within the molecules, rigid (use stored conformer only), rules (apply rules to accept or reject conformations during a conformational analysis), 2D key (calculate the hypothetical maximum and minimum distance between two features according to the connectivity). This key is used as a prefilter for 3D searching. The database may be rekeyed at any time, and the method chosen depends upon the search strategy as described below. At this stage the features are represented by a limited number of general definitions, hydrogen-bond donor, hydrogen-bond acceptor, aromatic center, and basic center, referred to as center types. (The distinction between different hydrogen-bond donors for example is achieved through the powerful atom-typing procedures available within Chem-X described in the next section.) The possible center−center distances are stored within one word (32 bit) screen sets, with each bit within the word representing a distance range bin;[15,16] all combinations of centers are stored, thus four centers would give ten word screens.

1.3. Atom-Typing. Effective 3D database searching requires that different atom environments can be clearly identified and distinguished from others. Important environments for pharmacophore searching are hydrogen-bond donors, hydrogen-bond acceptors, basic and acidic centers, aromatic rings, and hydrophobes. Methods have been developed in-house for the ChemDBS-3D system, based on the powerful atom/center perception possible in Chem-X using customized parameterization files and databases. It is important to be able to identify and distinguish both good and bad (deactivated) centers.

The method used in Chem-X uses a parameterization database with two types of special fragments. The atom types are defined automatically upon reading or building a 3D structure in Chem-X. The first type identifies environments via the number of bonds and the surrounding atoms; an atom must have the exact number of bonds to match, but bond order is disregarded. The second type identifies via the type (order) of bonds and the surrounding atoms ("*Z*" *type*); an atom must have the exact bond order matched and at least as many bonds (only the minimum number are defined). This type of fragment can be used to add "dummy" atoms for rings and lipophilic centers. The modifications made by the fragments are sequentially applied, a fragment can use modifications made by a previous fragment when identifying the environment; aromatic and double bonds are treated as equivalent, and addition of an aromatic ring centroid makes all the defining bonds aromatic. The logic used is to first modify for the general case and then modify special cases, putting the highest priority fragments last. Thus for benzamide, the nitrogen would first be identified as anilino type (i.e., attached to aromatic ring), and then the higher priority amide fragment would change it to an amide nitrogen; the first generic amide fragment would assign it the deactivated type (14 Q), and, if a hydrogen was attached, the second fragment would assign it the hydrogen-bond donor amide type (14E). Some possible atom types needed to differentiate nitrogen atom environments are shown in Table 2. Other atoms are attached to the atom to be modified in order to define its environment, and these can also be modified (e.g., both the C and N of C═NR) at the same time; atomic partial charges can also be set. An advantage of the atom-typing method is that assignment can be made by expected protonation state at physiological pH, for example, with bases (protonated; H-bond donors) and acids (deprotonated; H-bond acceptors), though these groups are read in as RNR_2 or RCO_2H; thus by atom type alone it is possible to identify the required type of group. A comparison of different methods used for defining atom environments was recently published[17].

Aromatic ring centroids are represented as dummy atoms added using the second type (Z) fragments in the parameterization database. Aromatic rings are identified by bond order, multiple fragments being needed as many rings do not have all bonds aromatic or double-single alternated, especially when heteroatoms are present (e.g., furan, thiophene, pyrrole, and imidazole).

Lipophilic/hydrophobic regions need to be identified and appropriate dummy atoms placed at their core. This can be achieved in a limited way through the use of predefined hydrophobic fragments but are best defined using the environment. We have studied three methods involving both fragments and environment for achieving this objective:

(i) A simple fragment approach in which a small collection of fragments are nominated as hydrophobic (i.e., isopropyl, butyls, cyclohexyl, etc.) and treated in a way analogous to the aromatic centers described above; however, it is not possible to cover all possibilities, and the method can become cumbersome.

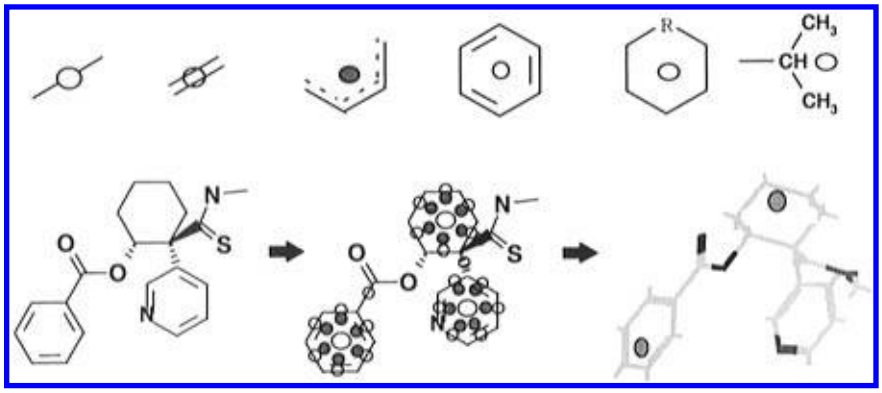


Figure 2. An example of the type of fragments used to add potential lipophilic atoms and their use.

Table 3. Pharmacophore Group Definitions

generic group type	comments
hydrogen-bond donor	basic/acidic groups excluded
hydrogen-bond acceptor	unsubstituted imidazole, both nitrogens included as donor and acceptor; basic/acidic groups excluded
hydrophobe	defined using hydrophobic substructures, e.g., *tert*-butyl and isopropyl groups, or as described in text
aromatic centroid	five- and six-membered aromatic rings
acid center	carboxylic and tetrazolic
basic center	aliphatic amines, amidines, guanidines, hydrazines

(ii) A bond polarity approach which became incorporated into the Chem-X package and is based on atom electronegativities and uses an algorithmic search for regions with nonpolar bonds. The electronegativity difference between two bonded atoms is calculated, and the atoms are removed from the hydrophobe list if the difference is greater than a predefined value, 0.45. All remaining atoms are potentially hydrophobic, a hydrophobic region being defined (using a "dummy" atom) if three or more hydrophobe atoms are bonded. If there are greater than eight such bonded atoms, attempts to split the region are made; however, multiple ring systems (e.g., steroids) currently may have only one center and are best identified by the environment, as described below.

(iii) An electrostatic environment method has been developed in-house using Chem-X and places dummy atoms at the centers of regions which are identified as nonpolar according to the local "electrostatic potential"; it is empirical and was designed to give an intuitive placement of such centers. Initially all possible potential lipophilic centers are identified using simple fragments (example fragments are illustrated in Figure 2) and represented by placement of a dummy atom. They are then evaluated in turn and eliminated as necessary according to the local electric field: slightly modified Gasteiger charges are placed on the heteroatoms, and a charge of +1 is placed on the dummy atom; the electrostatic energy between the dummy and all heteroatoms within 4.2 Å is then calculated, and the dummy is eliminated if it is too polar ($E < -28$ kcal/mol) or too close (2.35 Å) to another previously accepted lipophilic center. Unsubstituted isopropyl and butyl group dummy atoms are an exception to this elimination rule and are retained regardless of the calculated potential. An example structure is shown in Figure 2. The objective of this procedure is the automatic placement of hydrophobic centers in positions which usually would be selected by an experienced medicinal chemist.

In our experience (iii) is the method of choice, but the algorithm is rather slow when working with large databases. Method (ii) is fast, even for large databases, and produces good results for many compounds, but it does currently produce some anomalies in its center placements. Method (i), although not as thorough as the other two methods, is the quickest and simplest and was selected for the initial work reported in this paper (fragments: isopropyl, butyls, and cyclohexanes).

1.4. Pharmacophore Definitions and Query Definitions. 1.4.1. Pharmacophore Groups. Six pharmacophore groups were chosen because of their importance in receptor-ligand binding; these are listed in Table 3. Through the use of parameterization it is possible to distinguish easily acidic centers from other acceptors and, for example, to equivalence tetrazolic and carboxylic acids in searching by just selecting the two atom types. Tautomerism is treated by including certain atom types in both the donor and acceptor class, for example, unsubstituted imidazole nitrogen. Certain atom types with very weak properties were deliberately excluded, for example, thiols as donors. Special atom types have been created to distinguish strong, medium, and weak hydrogen bond donors and acceptors (e.g., *N*-oxide and nitro groups; carbonyl and ether oxygen of esters); oxygen atoms attached to other heteroatoms, for example, phosphate and sulfate oxygens are also distinguished and can be excluded if desired. Quaternary nitrogens were not included, and basic nitrogens were considered only as basic centers and not as general H-bond donors. The definition of a particular pharmacophore group can be easily modified simply by the inclusion or exclusion of an atom-type from the equivalence list.

Each "pharmacophore class" is defined by three atom/environment centers with the appropriate atom types associated with each atom (and an attached hydrogen if the atom is a donor). All nonredundant combinations of these center types are generated—56 in total.

1.4.2. Pharmacophores. Each pharmacophore is defined by a combination of center type and center−center distance. The center−center distance has a tolerance associated with it. A preliminary study of a subset (ca. 10 000 compounds) of the corporate database was performed to best define the distances and tolerances. Using the pharmacophore classes defined above, queries were generated for center−center

distances covering the range 3−24 Å in 3 Å steps (taking due account of symmetry and the triangle inequality—see below), i.e., distances 4.5 ± 1.5 Å, 7.5 ± 1.5 Å, etc. The queries were used to search the structures as stored in the database (no conformational analysis). A further set of searches, covering the distance range 2−3 Å for donor−acceptor distances, was performed. This sampled amides, for example, as two distinct pharmacophore centers. Analysis of the results allowed modified distance ranges to be defined, with smaller ranges for the shorter distance as these give greater numbers of hits. The final set of distances used in this work are as follows: 2−4.5, 4.5−7, 7−10, 10−14, 14−19, and 19−24 Å.

1.4.3. Query Generation. Queries, as defined by the pharmacophore definitions described above, were generated using scripts written in the Chem-X programming control language (PCL). Each combination of three-point pharmacophore and distance is generated with the following considerations:

(1) Symmetry—duplicate queries are not generated if a pharmacophore definition contains two or three identical pharmacophore groups.

(2) Triangle inequality—for certain distance combinations it is geometrically impossible to define the query using the midpoints of the distance ranges to specify the atom positions. However, it may be possible to generate a query covering only part of the required distance range. In this case the additional constraint is applied that the minimum allowed tolerance on a distance is 0.25 Å, for a query to be generated.

There is no direct mechanism within Chem-X for storing queries in a database. Thus a procedure was devised for storing the queries in the database indirectly. The 3D structure is written to the database and database fields used to store tolerances and lists of atom types. The use of a database allows the use of answer sets to search subsets of queries and storage of certain information at search time such as the total number of hits for each query.

1.5. Database Searching Strategy. A PCL script controls the search process. Each query is regenerated from the information stored in the query database and the structural database searched. ChemDBS-3D provides several options for searching, relating in particular to conformational regeneration. A fixed conformer search searches only the stored geometry. Rule-based searching applies conformational rules to accept or reject conformations during a conformational analysis, which can be either systematic or random. Flexifit applies a "torsional tweaking" algorithm. Both methods have the option of "bump-checking", to eliminate those conformations (or families of conformations) with atom contacts (as defined by VDW radii). For the search to be most efficient the database should be "keyed" appropriately: fixed conformer, rules, or 2D key. The keys indicate the presence or absence of a particular center−center distance (using the ranges described for ChemDiverse plus >15Å); the centers used are normally hydrogen-bond donors, hydrogen-bond acceptors, charged/basic centers, and aromatic ring centroids/hydrophobic centers. The fixed conformer key is based on the 3D distances present in the stored conformation, the "2D" key is from the lower and upper bounds of the 3D distances based on connectivity only, and the rule-based key is calculated from actual 3D center−center distances observed during a keying rule-based conformational analysis step. The latter can be a CPU intensive calculation, and each set of keys give most effective filtering with the complementary search method; the fixed conformation key is only suitable for the rigid search and flexifit needs 2D keys.

The choice of search strategy depends upon the size of the database and the accuracy required. As outlined here, the method is slow on very large databases, such as a large corporate database, unless fixed conformer searches are used. However, for smaller databases, correctly keyed, the time constraints are much less severe and full conformational searching can be used. Rule-based searching can be faster than flexifit and generally provides a reasonable geometry for a hit. Flexifit does generally give more hits, particularly with large distances and small tolerances, because it is not restricted to a defined set of torsion angles but can give high energy conformers, particularly if bump-checking is not used.

Chem-X will reject a solution that meets the distance constraints if the matching atoms do not fit onto the query atoms/centers within a defined tolerance (of RMS value for a superposition). This has been shown to be a very important refinement to 3D database searching.[9] In this work the fitting tolerance was defined as being equivalent to the largest distance tolerance within the query.

The results of the search with each query are stored within an answer set as a list of compounds which hit a particular query. For the results presented in this work the April 1994 or July 1994 versions of Chem-X were used.

1.6. Generation of Molecular Descriptor and Library Comparison. A series of UNIX scripts and C-programs have been written to collate the data stored in the answer sets. The output files detail for each query the total number of hits and the compounds matched. At a compound level the queries which match that compound are listed. This information is used to generate the molecular descriptor, which associates each pharmacophore with a bit, bits that are set "on" indicate that the compound contains the pharmacophore. It is thus possible to compare libraries at a pharmacophore level—which pharmacophores are contained within a library and how many occurrences—and at a compound level—how similar are compounds within a library or within different libraries according to the pharmacophores they contain?

Similarity values quoted refer to the Carbo Index[18,19] for comparisons involving frequency data or the Tanimoto coefficient[20] when two bit-strings are being compared.

2. DESCRIPTION OF CHEMDIVERSE STRATEGY

ChemDiverse is being developed by Chemical Design Ltd. and, like the PDQ strategy, generates a library or compound descriptor based upon pharmacophores expressed by the compounds in the library. However, ChemDiverse approaches the problem in a different manner. Conformational sampling (normally a single rule-based systematic or random conformational analysis) is performed for each compound in the database, and for each accepted conformation all hypothetical three-point pharmacophores are calculated and stored by setting the appropriate bits in the pharmacophore key. No attempt is made to fit the conformation onto the pharmacophore centers. The pharmacophore key is indexed according to all hypothetical three-point pharmacophores covering 30 center−center distances, by default 1.7−3.0 Å in 0.1 Å steps; 3.0−7.0 Å in 0.5 Å steps; and 7−15 Å in 1

Table 4. Results Obtained from Profiling Various Libraries with PDQ and ChemDiverse[a]

	PDQ		ChemDiverse	
library (no. of compds)	pharmacophores (max. 5916)	%	pharmacophores (max. 160 844)	%
corporate DB (150 000)	5705[b]	96.4	157 015	97.6
fingerprint subset (11 000)	5470[c]	92.5	126 740	78.8
ACD (105 000)			147 304	91.6
benzodiazepine library (1232)	1366[c]	23.1	12 186	7.6
LIB1 (1678)	2391[d]	40.4		
LIB2 (900)	1285[d]	21.7		

[a] For definitions of libraries see text. All ChemDiverse analyses were performed with rule-based conformational analysis (version with pharmacophores derived from four different center types). For PDQ the analysis was performed as follows. [b] Rigid geometry. [c] Flexible fitting. [d] Rule-based searching.

Å steps; these can be customized to give a larger distance range with larger increments. For the implementation used in this work (July 1995 version), the three centers are chosen from four generic types: hydrogen-bond donors, hydrogen-bond acceptors, positively charged/basic nitrogen; and aromatic ring/hydrophobic groups. The 20 possible center combinations with the 30 distance ranges gives a total of 160 844 valid three center pharmacophores; the 540 000 theoretical combinations are reduced by geometric (triangular inequality) and symmetry considerations. A new implementation, available from the October 1995 release, has an extension to seven center types for creating the three-point pharmacophores. This enables hydrophobes, acids, and bases to be considered separately, giving a total of 848 925 valid three center pharmacophores (from the combination of 84 three-center pharmacophores and 30 distances).

The atom type definitions have been extensively redefined using our in-house parameterization database, as described in the PDQ atom-typing section above. The ChemDiverse procedure is fast, running at greater than 1000 compounds per hour on a Silicon Graphics workstation (R4400−200 Mhz) with corporate database compounds and requires only one complete search of the database; very flexible compounds (>10 rotatable bonds) can be much slower, and random rule-based sampling can be more efficient. However, the information obtained is limited in this version to a pharmacophore key, normally calculated for the whole database or subset thereof. An individual molecular descriptor is not normally output or saved, although by using small subsets (e.g., one compound) this can be done. The key however can only currently be saved in relatively large individual files, including space for the inaccessible/unmatched pharmacophores. [Chem-X versions from July 1996 produce more efficient size key files, and functionality was introduced to count the number of times a pharmacophore is matched.] By applying logical comparisons of this key at search time to others of interest (e.g., calculated from active ligands or complementary to an active-site) and outputting the number of overlapping pharmacophores a powerful form of 3D multipharmacophoric similarity searching can be performed. This key marks only the presence or absence of a pharmacophore, and no count of the number of times a particular pharmacophore is hit is obtained. A full range of logical operations can be performed on different keys within Chem-X/ChemDiverse.

3. RESULTS

3.1. Profiling Large Databases. A database of 150 000 structures from the RPR corporate database has been analyzed using the PDQ and ChemDiverse methods. Results are shown in Table 4, together with those for some other databases ("Fingerprint subset": 11 000 "diverse" compounds selected from the RPR database from Daylight fingerprint clustering, see below; "ACD": Available Chemicals Directory;[21] "Benzodiazepine library": virtual combinatorial library, see 3.2; LIB1 and LIB2: in-house combinatorial libraries, see 3.2). For PDQ analysis of the corporate database, fixed geometry searching was used initially for all the queries to enable an acceptable search time; the goal of the study was to know which pharmacophores were matched in the database, and it was expected that many would be found without needing conformational regeneration. However, the 211 queries with no hits found in the RPR database were used to search the database again, this time with conformational flexibility, and all but 10 pharmacophores were matched by structures in the database. Whilst the use of a mainly fixed conformer search with PDQ would be a limitation for a thorough analysis (i.e., to generate a detailed molecular descriptor for each molecule), it can be adequate where the goal is simply to determine the pharmacophoric diversity of the database.

With both methods it can be seen that the database covers most of the available pharmacophores. Figure 3a,b represent further results from the PDQ analysis and shows the utility of having quantitative information about the queries and the molecules. Figure 3a shows that the number of hits falls off as the center−center distances increase. It can be seen from Figure 3b that the database contains a number of compounds with no three center pharmacophores and a number with more than 100.

A subset of compounds had been selected previously from the corporate database utilizing Jarvis−Patrick clustering on Daylight 2D structural fingerprints.[22] This set of 11 000 compounds has been analyzed for pharmacophore diversity. For ChemDiverse this entailed a reanalysis of this subset of the database. For PDQ a list of satisfied pharmacophores for the stored conformation was obtained from the detailed output files of the analysis of the corporate database. The 1550 unmatched pharmacophores were used to search the subset of compounds using the flexible fitting options in ChemDBS-3D. With ChemDiverse (using pharmacophores from four center types) the set covers approximately 78% of the theoretical pharmacophore space, about 80% of the possible pharmacophores for the whole database. The set covers about 92% of the space defined in the PDQ analysis (about 95% of the pharmacophores found for the whole database).

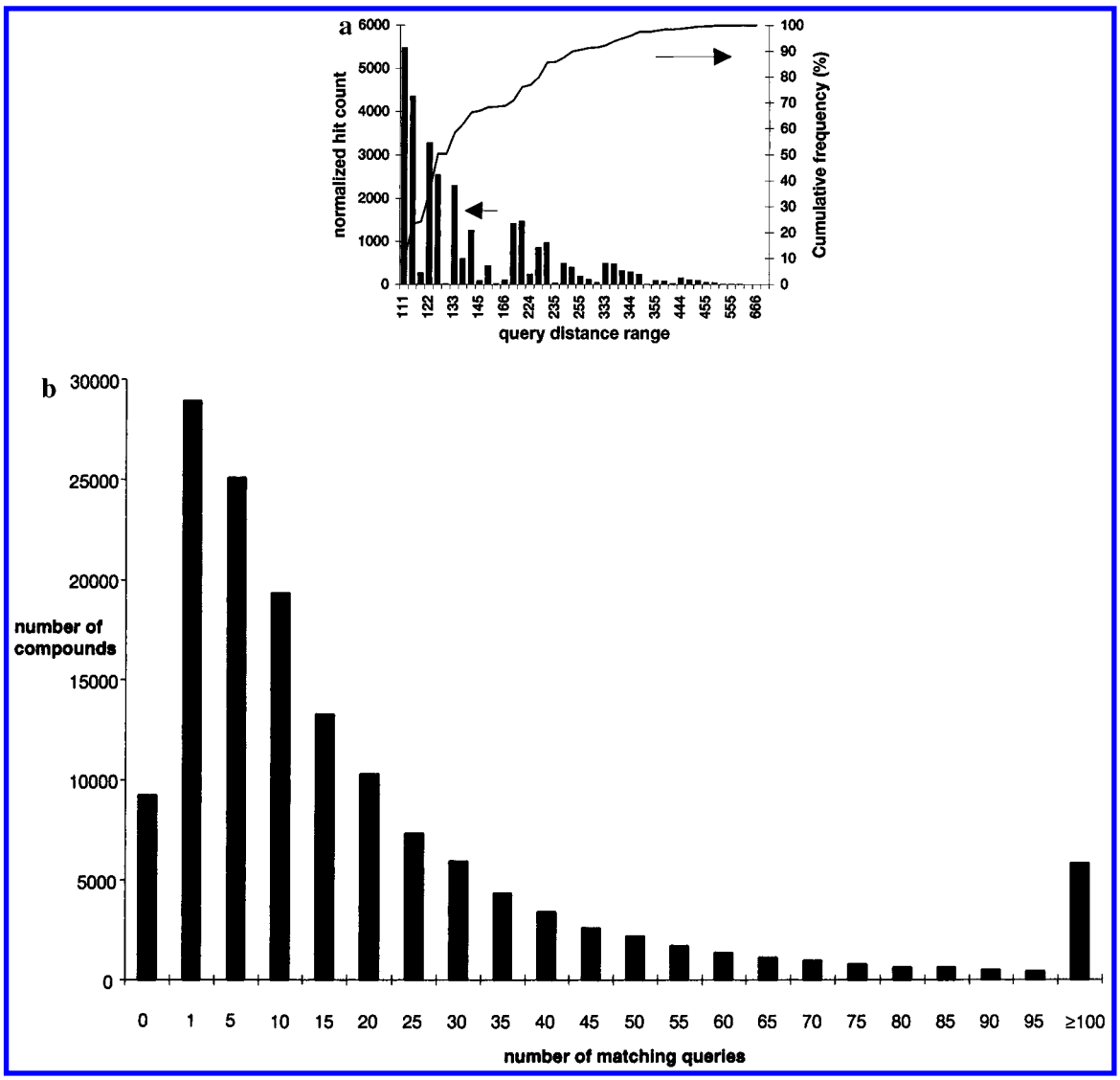
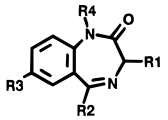


Figure 3. (a) Results from the analysis of 150 000 compounds from the corporate database using the pharmacophore-derived queries approach. Distribution of number of compounds matching a pharmacophore as a function of pharmacophore distance. The combination of shortest distance is 111; 666 is the longest distance. (b) Results from the analysis of 150 000 compounds from the corporate database using the pharmacophore-derived queries approach. Histogram of number of matched pharmacophores per compound.



R1 from	Gly, Ala, Val, Phe, Trp, Asp, Asn, Glu, Gln, Thr, Lys	[11]
R2 =	Ph, 4-MeOPh, c-Hex, 2-thienyl, 4-NO_2Ph, 4-NH_2Ph	[6]
R3 =	H, Cl, NO_2, OMe, NH_2	[5]
R4 =	H, Me, benzyl, i-Bu	[4]

Figure 4. Description of the hypothetical benzodiazepine combinatorial library used, an expanded version of the library of De Witt et al.[23]

3.2. Profiling Combinatorial Libraries. An expanded/modified version of the benzodiazepine library proposed by De Witt et al.[23] has been used to test the PDQ and ChemDiverse strategies. The virtual library consisted of 1232 structures as shown in Figure 4. The pharmacophore profiling results are shown in Table 4. The PDQ and ChemDiverse results were generated with rule-based conformational searching/analysis. The analysis showed that the library was particularly deficient in pharmacophores involving basic groups. This reflects to some extent the choice of R-groups and is a characteristic of the library. The molecular descriptor generated by the PDQ approach allows a comparison of the molecules in the database. The average maximum similarity of a compound to all other compounds in the library is 0.94. This is high and is in part due to the functional groups present in the core of the molecules which will contribute a significant number of pharmacophores to all the structures.

Chemical Design have implemented a procedure with ChemDiverse/Chem-X that enables compounds to be selected from a library in such a way that the pharmacophore space is covered most efficiently. The compounds are first "diverse" sorted, by default using the class, bond, and "2D" center−center distance keys as a similarity measure; the first compound is the one nearest the mean of the keys, the next the furthest away, and the next ones the furthest from those already selected. This sorted set is then analyzed for pharmacophores, with full conformational sampling (e.g., rule-based systematic conformational analysis), compounds only being retained if their pharmacophore key is dissimilar enough to the key for the compounds already selected. The process is repeated until a defined proportion of the total pharmacophore space of the library is covered. A compound is also rejected if it is too flexible or too rigid (as defined

by the user using the number of rotatable bonds). The criteria for dissimilarity for a compound to be added to the list is a maximum predefined percentage of pharmacophore key overlap with previously selected compounds; 90% is a default value, but several consecutive analyses can be made, gradually increasing the acceptable overlap. Such a procedure was carried out on the benzodiazepine library. With just 7% (93) of the 1232 compounds it is possible to cover over 90% of the pharmacophore space of the library, increasing to 9.6% covering 96.1% of the possible pharmacophores upon further analysis (95% overlap cut-off).

This result further highlights the redundancy of the compounds in covering pharmacophore space, as shown by the PDQ similarity calculations. Such information is of use to the library designer, in conjunction with the more detailed molecule profiles as shown for the corporate database, to refine the choice of substituents if so desired. The PDQ molecular descriptor can also be used to compare molecules between libraries. This is important as two libraries might cover similar pharmacophore space but in a different manner at the molecule level. Thus the libraries can be selected accordingly. As an example, the two in-house libraries shown in Table 4 that were analyzed by PDQ using rule-based conformational analysis can be compared. LIB2 (900 compounds, 1285 pharmacophores) covers a subset of the pharmacophores of LIB1 (1678 compounds, 2391 pharmacophores), with an overlap of 1060 queries (similarity 0.6 at the query level); however, when the distribution of molecules satisfying the various pharmacophores is taken into account (i.e., comparing not only the presence of a pharmacophore in both libraries but how frequently it is covered) it can be seen that the difference between the two libraries becomes quite large, with the similarity dropping to 0.3 (similarity is assessed using the Carbo Index). Also, by calculating the library self-similarity, a tendency for compounds in LIB2 to be more diverse from each other than those in LIB1 is shown (LIB1 self-similarity 0.85; LIB2 self-similarity 0.64).

4. DISCUSSION

4.1. 3D Pharmacophores as a Means of Assessing Diversity. The method of pharmacophore-derived queries (PDQ) described in this paper provides a means for assessing the diversity of compound libraries, utilizing 3D structural information and taking account of conformational flexibility. The ChemDiverse approach, being developed by Chemical Design is similar in philosophy but is implemented differently. The two approaches are complementary. PDQ provides detailed molecular information coupled with pharmacophore and whole library data, whilst ChemDiverse has the advantage of speed and finer resolution in the pharmacophore distances.

The results obtained with profiling subsets of compounds selected by other methods shows the importance of utilizing 3D information. For example the 11 000 compound subset of the corporate database, selected using a Jarvis−Patrick clustering of Daylight fingerprints, provides a reasonable coverage of the pharmacophore space, about 80% coverage with just 7% of the compounds compared to the corporate database. In one regard this is not surprising as the functional groups used to define the pharmacophores are also included in the 2D structural characterization. On the other hand, there is still about 20% of pharmacophore space observed in the whole database to be covered if the subset is to be truly representative of the pharmacophoric (three-point) diversity of the database (see Table 4).

The above discussion raises a second issue. That is, how well does a compound satisfy the pharmacophore. The pharmacophore is considered as a necessary condition for a compound to be biologically active but is not sufficient on its own to ensure activity; other 3D properties of the molecule such as the detailed surface shape will be important. In the work described here both methods utilize only the three pharmacophore points in assessing the quality of a fit, but there are two simple concepts that can be used to improve the quality of the fits selected: pharmacophore accessibility and pharmacophore ratio. The first involves a check that the pharmacophore centers matched to a query are actually accessible for a receptor interaction, whereas the second is to check that the size of the pharmacophore corresponds to a reasonable proportion of the molecule. Recent developments in the ChemDiverse methodology have given rise to very quick approximate solutions for both. Pharmacophore accessibility is gauged by estimating the direction of the interaction for hydrogen-bond donor and acceptor centers on the molecule and rejecting if the vector points inwards (i.e., within the triangle defining the three-point pharmacophore). Pharmacophore ratio is calculated by comparing the area of the three-point pharmacophore (becoming a volume with four or more points) with the heavy atom count for the molecule; preferred solutions have a large ratio (i.e., the majority of the molecule is involved in the presentation of the pharmacophore groups). The nature of the ChemDiverse approach means that these methods are necessarily approximate, but they are rapid. However, in the PDQ approach the pharmacophore is generated as a 3D query and used to search the database. This gives access to a wide range of possibilities with regard to assessing the shape of molecules matching a pharmacophore and the accessibility of pharmacophore groups, although final searching speeds with full conformational sampling are slower. Methods to improve the descriptor by inclusion of a more detailed shape component than is available through the use of three pharmacophore points are under development; this area will be the subject of a subsequent paper. The size issue could be dealt with quite simply by excluding compounds which exhibit pharmacophores with large distances from the match lists of compounds exhibiting pharmacophores with exclusively short distances. Any compounds selected using this approach would then be more ideal candidates for a particular pharmacophore. Both methods identify "promiscuous" compounds, those that match a large number of pharmacophores, enabling the pharmacophores from these compounds to be treated separately; this information can be used to try to minimize such potentially flexible nonselective compounds in library design or compound selection.

4.2. Designing Chemical/Combinatorial Libraries Using Pharmacophore Diversity. The PDQ approach provides much useful information for profiling and comparing compound libraries. A histogram is obtained for the pharmacophores matched; this is important as a simple on/off bit string loses the count information. When comparing two databases it is important to consider how many times each pharmacophore is hit, as shown by the examples given above.

This full profile at both a pharmacophore and molecule level presents the library designer with a large amount of information which can be utilized in refining the choice of substituents and hence the pharmacophore distribution of compounds in the library.

The library designer must make a number of choices. What portion of pharmacophore space should be covered? How many times should a pharmacophore be covered? The discussion on shape in the previous section would suggest that a pharmacophore should be covered more than once in a library. The promiscuity of individual compounds within the library is important. Compounds which hit a large number of pharmacophores will either tend to be very flexible or have many functional groups which could lower the binding compared to a compound exhibiting a particular pharmacophore more succinctly. Specificity can also be a problem, such as where a mainly hydrophobic pharmacophore is in a molecule with many charged groups or a small pharmacophore is part of a much larger molecule; atom type (e.g., basic, acidic) specificity could of course be a major problem, but adequate solutions are available (see parameterization section). Such information is available with the PDQ method described here and, currently without the pharmacophore count, within the ChemDiverse method. In addition it is possible to compare the library to previously synthesized or other hypothetical libraries to assess the added value of the library in a pharmacophoric sense and to identify missing diversity. This latter point is an important benefit of the systematic analysis of pharmacophoric space; clustering methods do not enable such missing diversity to be identified, and the comparison of different libraries usually requires reclustering. As well as being able to identify and design these diverse general libraries, it is possible to identify and design biased and focused libraries. A biased library can be designed to be enriched in certain desired pharmacophores (e.g., from other active ligands or from target site information). A focused library can build upon structural restraints to explore other regions in a diverse and/or enrichment mode.

4.3. Comparison of PDQ and ChemDiverse. Both PDQ and ChemDiverse are based on sampling pharmacophore space, though they approach the problem from different directions. ChemDiverse analyzes the database once, with conformational regeneration; this is normally done using a systematic rule-based conformational analysis, but random or other sampling methods can also be used, such as with very flexible molecules. The information provided is essentially the pharmacophore coverage of the library, either whole or in subsets, at a very fine level (small distance ranges per pharmacophore) but with only limited individual molecule information easily being obtainable at present. With PDQ a smaller number of pharmacophores are generated (larger distance ranges) and used to individually search the database; this is slower but gives more detailed information on a per molecule basis, allowing comparisons between libraries at both the molecule and library level. Intralibrary comparisons are also possible—how similar are the compounds within the library? It is also possible to increase the speed of PDQ if all that is required is a basic profile of the pharmacophores present in the library, by the simple expedient of stopping each search once one hit has been found for a pharmacophore. This drastically reduces the CPU overhead by reducing the time spent searching queries with a large number of hits, but limits the per molecule information.

The ChemDiverse pharmacophore descriptor covers many more hypothetical pharmacophores than the PDQ approach because of the increased resolution of the distance bins and requires only one full conformational sampling; more work is required to identify the best combination of pharmacophore groups and distance ranges. The PDQ approach has adopted distance ranges/tolerances and pharmacophore group definitions which we have used frequently in-house for 3D searching. It uses separate searches and thus is slower as conformational analysis may be performed several times on a molecule. Given the limited resolution of torsion space within normal rule-based conformational analysis, some of the additional information available through the higher resolution of the ChemDiverse descriptor may not be significant; a modification of the default distance range and increments is possible and likely to be beneficial. The increase in the number of possible pharmacophore classes in ChemDiverse from four to seven is an important development, in particular the separation of acids and of hydrophobes/lipophiles.

The two methods have complementary features. ChemDiverse provides a relatively quick method for obtaining pharmacophore profiles, often only at a library level, using very large numbers of pharmacophores. The PDQ approach is slower, with a much more limited number of pharmacophores feasible, but provides more detailed information, particularly at the molecule level, allowing comparisons at both a library and molecule level. The fact that PDQ generates the pharmacophore and searches the database, producing a database of hits superimposed on each pharmacophore query, provides the opportunity for a much more detailed analysis, as noted above for shape analysis.

CONCLUSIONS

The work presented in this paper has shown how it is possible to utilize 3D information to profile compound libraries and aid the design of chemical/combinatorial libraries. Both the PDQ and ChemDiverse methods use the concept of 3D pharmacophores to describe the compounds, based on 3D distances between generic features believed to be important for ligand−receptor interactions; both currently use three center pharmacophores but need not be limited to this. The in-house PDQ method described in this paper provides detailed information of the pharmacophores covered and how the compounds in the library cover those pharmacophores. This information is of value to the library designer and enables comparisons to be made between different libraries, using pharmacophoric similarity as the similarity measure. The ChemDiverse method that has recently become commercially available enables a much faster profiling of a finer resolved pharmacophoric space but currently with less information readily accessible at the individual compound or pharmacophore level. Both methods enable different sets of compounds (e.g., libraries) to be compared, and missing pharmacophoric diversity or desired pharmacophoric similarity to be identified.

The methods have a wide range of applications beyond designing and profiling combinatorial libraries. The pharmacophore information can be used to partition a large database into a diverse subset covering pharmacophore space.

If several diverse leads are identified during screening, then the methods provide a method for pharmacophore identification by ANDing the molecular descriptor. The descriptor may also be used in similarity searching of leads against a collection of compounds. This would be useful in the situation that arises frequently, where it is not possible to identify the key pharmacophores, and would allow the identification of molecules with similar pharmacophore profiles, thus giving more detailed structure−activity relationship information (SAR) at an early stage of a project. This area is being actively pursued within the group.

ACKNOWLEDGMENT

The authors thank R. A. Lewis for valued discussions.

REFERENCES AND NOTES

(1) Presented in a preliminary form at the MGMS Meeting, York, April 1995; see also ref 7.
(2) Downs, G. M.; Willett, P. Similarity searching and clustering of chemical-structure databases using molecular property data. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1094−1102.
(3) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring diversity: Experimental design of combinatorial libraries for drug discovery. *J. Med. Chem.* **1995**, *38*, 1431−1436.
(4) Mason, J. S.; McLay, I. M.; Lewis, R. A. Applications of Computer-Aided Drug Design Techniques to Lead Generation. In *New Perspectives in Drug Design*; Dean, P. M., Jolles, G., Newton, C. G., Eds.; Academic Press: London, 1995; Chapter 12, pp 225−253.
(5) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical similarity using physicochemical property descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.
(6) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical similarity using geometric atom pair descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128−136.
(7) Ashton, M. J.; Jaye, M. C.; Mason, J. S. New Perspectives in Lead Generation II: Evaluating Molecular Diversity. *Drug Discovery Today* **1996**, *2*, 71−78.
(8) ChemDBS-3D/ ChemDiverse: developed and distributed as part of the Chem-X modeling package by Chemical Design Ltd., Roundway House, Cromwell Park, Chipping Norton, Oxfordshire, OX7 5SR, UK.
(9) Greene, J.; Kahn, S.; Savoj, H.; Sprague, P.; Teig, S. Chemical function queries for 3D database searching. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1297−1308.
(10) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.
(11) 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.
(12) Lewis, R. A., personal communication.
(13) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, 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.
(14) Pearlman, R. S. *CDA News* **1987**, *2*, 1−7. Rusinko, A., III; Skell, J. M.; Balducci, R.; McGarity, C. M.; Pearlman, R. S. University of Texas, Austin. CONCORD, A Program for the Rapid Generation of High Quality Approximate 3D Molecular Structures. Distributed by Tripos Inc., 1699 Hanley Road, Suite 303, St. Louis, MO 63144 U.S.A.
(15) Murrall, N. W.; Davies, E. K. Conformational Freedom in 3D Databases. 1. Techniques. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 312.
(16) Good, A. C.; Mason, J. S. Three-Dimensional Structure Database Searches. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1996; Vol. 7, Chapter 2, pp 67−117.
(17) Mason, J. S. Experiences with Searching for Molecular Similarity in Flexible 3D Databases. In *Molecular Similarity in Drug Design*; Dean, P. M., Ed.; Blackie Academic and Professional: Glasgow, 1995; Chapter 6, pp 138−162.
(18) Carbo, R.; Leyda, L.; Arnau, M. An electron density measure of the similarity between two compounds. *Int. J. Quantum Chem.* **1980**, *17*, 1185−1189.
(19) Carbo, R.; Domingo, L. LCAO-MO similarity measures and taxonomy. *Int. J. Quantum Chem.* **1980**, *17*, 1185−1189.
(20) Johnson, M. A. A review and examination of the mathematical spaces underlying molecular similarity analysis. *J. Math. Chem.* **1989**, *3*, 117−145.
(21) Distributed by MDL Information Systems Inc., 14600 Catalina Street, San Leandro, CA 94577, U.S.A.
(22) Daylight User Manual, release 4.41, Daylight Chemical Information Systems Inc., 18500 Von Karman Avenue, Suite 450, Irvine, CA 92715 and Santa Fe, NM. Lewis, R. A.; Mason, J. S.; Menard, P. R., manuscript in preparation.
(23) DeWitt, S. H.; Kiely, J. S.; Stankovic, C. J.; Schroeder, M. C.; Cody, D. M. R.; Pavia, M. R. "Diversomers": An approach to nonpeptide, nonoligomeric chemical diversity. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 6909−6913.

CI960039G