# Bioisosterism as a Molecular Diversity Descriptor: Steric Fields of Single "Topomeric" Conformers

Richard D. Cramer,* Robert D. Clark, David E. Patterson, and Allan M. Ferguson

*Tripos Inc., 1699 South Hanley Road, St. Louis, Missouri 63144*

*Received April 18, 1996*[⊗]

The comparative molecular field analysis steric field of a single "topomeric" conformer is introduced as a molecular diversity descriptor particularly useful for combinatorial chemistry involving variations around a fixed "core". Using this new descriptor, 736 commercially available thiols are divided into 231 bioisosteric clusters, whose compositions agree at least as well with medicinal chemical experience and intuition as do clusters derived from Tanimoto differences between 2D fragment occurrences. However, in practice topomeric steric fields complement 2D fingerprints, being the two most frequently useful descriptors yet found for neighborhood-based design of combinatorial libraries.

## Introduction

Experimental techniques for very efficiently preparing and screening combinatorial mixtures or arrays of organic compounds are being developed and deployed at a remarkably high rate throughout the pharmaceutical industry. Even with a throughput of thousands of compounds per day, however, the numbers of compounds synthetically accessible exceed by many orders of magnitude the numbers which can actually be made and tested.[1,2] Thus a need is emerging to select thousands of structures per week from astronomical numbers of candidates, either for discovering new leads (by maximizing the "molecular diversity" of candidate compounds, to avoid redundant testing) or for rapidly following up possible leads (minimizing "molecular diversity" by maximizing the "molecular similarity" of candidate compounds to the potential lead structure).

Of course, ideas of molecular diversity and similarity[3,4] are hardly new to medicinal chemists. Bioisosterism, the replacement of one fragment by another that is structurally dissimilar but similar in 3D shape (such as replacing phenyl by thienyl), is a long-standing strategy of molecular modification.[5]

Despite the assumed primacy of 3D structure, the emerging literature on molecular diversity has emphasized 2D measurements. One reason is that rough general associations between fragments and biological properties were validated statistically decades ago.[6] Another is convenience; all commercial molecular database software uses 2D fragment keys or "fingerprints" to expedite structural retrieval. But we believe that the most important reason is the challenge of formulating and validating appropriate 3D measurements. The situation resembles that in the field of quantitative structure−activity relationships (QSAR) 10 years ago, where 2D-based descriptors were used exclusively despite a general awareness that 3D-based descriptors would be more relevant to selective biological activity. Indeed, when an appropriate 3D QSAR technique, comparative molecular field analysis (CoMFA),[7] became available, its widespread acceptance[8] and application[9] support the expected importance of 3D descriptors in general. This paper explores the possibility that a 3D field-based approach like CoMFA might usefully characterize molecular diversity and similarity.

The usual challenge in applying CoMFA is the molecular alignment procedure. This is hard enough when analyzing a set of biological data believed to come from the same receptor. How is one to "align" molecules with respect to arbitrary and unknown receptors? Our response is based on the belief that CoMFA alignment efforts may often overemphasize a search for "receptor-bound", "minimum energy", or "field-fit" conformations. When congenericity exists, a useful alignment could result from simply overlaying the atoms that lie within some selected common substructure and arranging the other atoms according to general canonical rules, with any resulting steric collisions ignored.

This "topomeric" alignment procedure seems especially applicable to the design of a combinatorial chemistry library. Typically there is a central "core", such as $-SO_2NH-$, $H(NCH_2CO)_nOH$ (peptoid), or benzodiazepine, with variable fragments, at the open valences, being contributed by reactants of a particular class. Within the combinatorial products, this central core tethers each of the side chains contributed by any set of reactants into the same relative position in space. In the language of CoMFA alignments, the variable part of each reactant can thus be oriented by overlapping the bond that attaches the side chain to a central core and using a "topomeric" protocol to select a representative conformation of the side chain.

Assessing molecular similarity by direct comparison of steric volumes has been worked on by many other investigators, starting in 1980 with Carbo[10] and Hopfinger.[11] For the most part, these studies[12] have focused on methodological details of shape comparisons or on using shape differences within particular QSAR studies. In the following work, the central issue is instead whether this generalized "topomeric" procedure for modeling side chains, unrelated to any particular "active conformation", affords a generally useful pattern of shape differences.

The general idea of selecting diverse aromatic substituents from cluster analysis, based on other substituent properties, was first published by Hansch and Unger in 1973,[13] and van de Waterbeemd[14] was one of the first to cluster simple familiar aromatic substituents based on CoMFA fields. Some other diversity measurements

* To whom correspondence should be addressed.
[⊗] Abstract published in *Advance ACS Abstracts,* July 15, 1996.

S0022-2623(96)00291-9 CCC: $12.00 

based on 3D shape have recently been presented for use in library design. As one aspect of "peptoid" design, Martin *et al.*[15] have explored a "polyomino" approach, in which a reagent is classified based on the kinds of assemblies of cubes into which it can be made to fit. In the approach most similar to ours, Chapman *et al.*[16] applied the compass 3D QSAR methodology[17] to the selection of 20 maximally diverse carboxylic acids, based on seeking the maximally diverse alignment of each of the 2000 acids considered, as contrasted with the topomeric alignment described below.

Pharmacophoric pattern matching[18] is the basis for another general approach to 3D molecular diversity descriptors. Indeed, one group[19] advocates a somewhat simplified set of pharmacophoric triplets as a central organizing principle for combinatorial libraries, and another[20] opines that careful attention to this principle will allow construction of a "universal screening library" comprising just 20 000 compounds. On the other hand, in an exemplary validation study, Brown, Bures, and Martin[21] reported that 3D "fingerprints", collections of fragments defined by pharmacophoric pairs of atoms, perform rather worse than collections of 2D fragments in defining clusters that separate active from inactive compounds. We believe that any molecular diversity measurement should be validated empirically before being recommended for use in library design.

In an accompanying paper,[22] we present the "neighborhood plot" as a general method for validating diversity descriptors and show that a "topomeric steric field" diversity measurement and a variant of the widely used Tanimoto coefficient between "2D fingerprints" are the most frequently useful of 11 diversity descriptors considered. Below we describe the topomeric alignment method and apply it to a set of 736 commercially available thiols. Clustering only the steric CoMFA fields of the resulting conformers produced an intuitively very satisfying division of these compounds into 231 "bioisosteric" clusters, differing from each other in steric size by at least a $-CH_2-$ group. Indeed, the structural uniformity within clusters was so great as to suggest distinctive structurally-based names for each of these clusters. These findings extend and support the accompanying neighborhood plot validation results in highlighting topomeric steric fields as a particularly useful descriptor of molecular diversity.

## Methods

All work was done within SYBYL 6.2, enhanced with both SPL and C codes to perform some of the specialized calculations described below. The individual steps required less than 0.5 h each to process the 736 molecules on an SGI R4000 workstation with ample memory.

**Generation of Conformations.** Usually the modeler seeks low-energy conformations. However, if clustering on steric fields is to produce useful results, then the major goal in conformer generation must be that molecules having similar topologies should produce similar fields. To this end, the following simple topologically-based rules have been devised for generating a single, consistent, unambiguous, aligned "topomeric" conformation for any molecule lacking prochiral atoms.[23]

The starting point for alignment of a molecule is a Concord[24] model, which is then FIT (SYBYL command) as a rigid body onto a template 3D model by least-squares minimization of the distances between structurally corresponding atoms. By convention, the template model is originally oriented so that one of its atoms is at the Cartesian origin, a second lies along the *x*-axis, and a third lies in the *xy* plane.

Torsions are to be adjusted for all bonds which (1) are single and acyclic, (2) connect polyvalent atoms, and (3) do not connect atoms that are polyvalent within the template model structure (*i.e.*, do not change any geometry that was fitted to the template!). Unambiguous specification of a torsional angle about a bond requires a *direction* along that "torsional bond" and two *attached atoms*, one atom attached to each of the two atoms which define the bond. In defining the topomeric conformation, the *direction* is always taken to be "pointing away from the FIT atoms", a well-defined concept for acyclic bonds. The following precedence rules then determine the two *attached atoms*. From each candidate attached atom, begin growing a "path", layer by layer, including all branches but ending whenever another path is encountered (occurrence of ring closure). At the torsionally bonded atom that is closer to the FIT atoms, first the attached atom which begins the shortest path to any FIT atom is chosen. If there is more than one such atom, the next choice is the attached atom with the lowest *x* coordinate (coordinate values differing by less than 0.1 Å are considered to be equivalent). If a tie still exists, choice then goes to the lowest *y* and then to the lowest *z* coordinate. At the torsionally bonded atom *farther* from the FIT atoms, the first choice is an attached atom beginning a path that contains any ring. When other than one path contains a ring, subsequent choices are the atom whose path includes the most atoms and then the atom whose path includes the highest sum of atomic weights and finally the atom with the highest *x*, *y*, and then *z* coordinates.

The new setting of the torsional value depends only on whether the two bonds between the chosen attached atoms and the torsionally bonded atoms are cyclic or not. If neither are cyclic, the setting is 180°; if one is cyclic, the setting is 90°; and if both are cyclic, the setting is 60°. Any steric clashes that may result from these settings are ignored. In this study about 5% of the resulting conformers had at least one nonbonded distance that was less than 70% of the sum of their van der Waals radii.

As an illustrative example, consider generation of the topomeric conformer for the side chain shown at the top of Figure 1, in which atom 1 is attached to some core structure by the upper leftmost bond. Assuming that the alignment template for this fragment involves atom 1 only, *the bonds* whose torsions require adjustment are those connecting atom pairs 1,3; 1,2; 5,8; and 10,14. (Adding atom 3 to the alignment template would make atom 1 "polyvalent within the template model structure", so that the 1,3 bond torsion would then not be alterable.) The *direction* of these bonds (which immediately determines which part of the side chain moves as a torsion changes) points, as given above, away from the aligning template. In contrast, if a torsional change were applied to the "14,10" bond instead of the "10,14" bond as shown, *all* of the molecules except atoms 10, 14, and 15 (and 13 by symmetry) would move. To define a torsional change, atoms attached to each of the torsionally bonded atoms must also be specified. For example, setting torsion "5−8 to 60°" could yield four different conformers depending on whether it is the 6−5−8−13, 4−5−8−9, 6−5−8−13, or 6−5−8−9 dihedral angle which becomes 60°. To make such a choice, "paths" are grown from each of the candidate atoms in "layers", each layer consisting of all previously unvisited atoms attached to any existing atom in any path. In choosing among the four attached-atom possibilities of the 5,8 bond, the middle panel of Figure 1 shows the four paths after the first layer of each is grown, and the bottom panel shows the final paths. In the bottom panel, notice within the rings that not only is the bond between 3 and 7 not crossed but also atom 11 is not visited because the third layer seeks to include atom 11 from two paths, so both fail. The attached atoms chosen for the torsion definition become the ones that root the highest ranking paths, according to the above rules. For example, in Figure 1, attached atom 4 outranks 6 because its path is the only one reaching the core, and attached atom 9 outranks 13 because its path has more atoms, so that it is the 4−5−8−9 torsion which is set to a prescribed value. For the same reasons, the

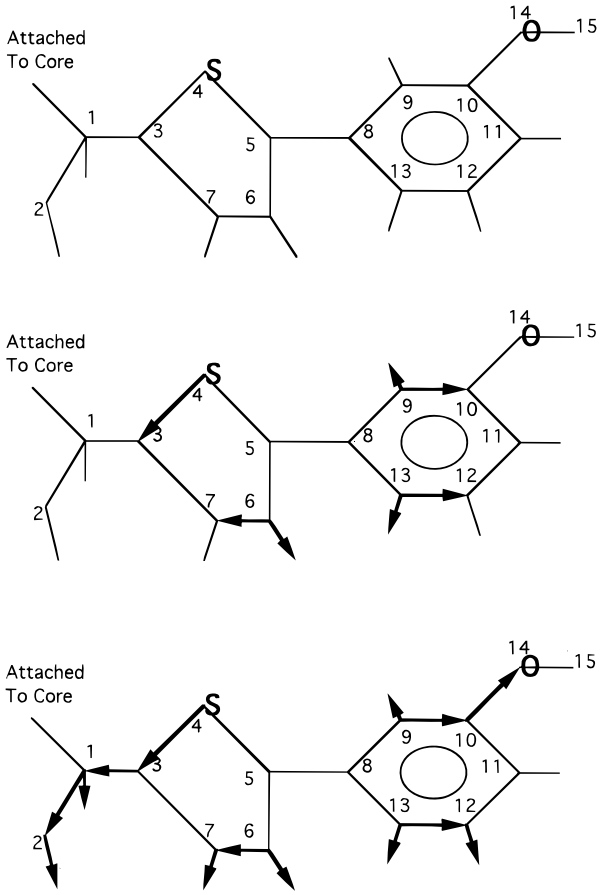



**Figure 1.** Example of the generation of a topomeric conformer. See text for details.

other complete torsions become 9−10−14−15, attached−1−3−4, and attached−1−2−16. The other sorting rules would be needed if atom 9 became aromatic nitrogen, with its hydrogen disappearing so that the 9 and 13 paths have the same number of atoms. Then the 9 path would still be chosen, since it has the higher molecular weight. If instead atom 14 is deleted, so that the 9 and 13 paths are topologically identical, the 9 path again is chosen because atom 9 has the same *x* coordinate (as drawn) but a larger *y* coordinate than does atom 13. As for the dihedral angle values themselves, torsion 1−5−8−9 is set to 60° because both the 1−5 and 8−9 bonds are cyclic; torsions 9−10−14−15 and attached−1−3−4 become 90° because only the 3−4 and 9−10 bonds, respectively, are cyclic; the attached−1−2−16 dihedral becomes 180° since neither the attached−1 nor the 2−16 bond is cyclic.

**Steric Fields.** The steric fields are generated almost exactly as in a CoMFA analysis, using an $sp^3$ carbon atom as the probe, on a 2 Å spaced cubic lattice measuring 18 Å per side, whose lowest *x*, *y*, *z* coordinates are −4, −12, −8 Å, with a cutoff (maximum value) of 30 kcal/mol for lattice points whose total steric interaction with all side-chain atom(s) is greater than the cutoff.

An important difference from the usual CoMFA field calculation procedure is that atoms which are separated from any template-matching atom by one or more rotatable bonds make reduced contributions to the overall steric field. The attenuation factor used in this work for each atom was $0.85^n$, where *n* is the number of rotatable bonds separating that atom from any FIT atom. Thus, in Figure 1, atom 1 has a weight of 1.0; its attached hydrogen and atoms 2 and 3 have weights of 0.85; atoms 4−7 with their attached hydrogens and the hydrogen attached to atom 2 have weights of $(0.85)^2$ or 0.72, and so forth.

**Cluster Analysis.** All classifications described were obtained by "complete linkage" hierarchical cluster analysis of the resulting steric field matrices, using "CoMFA_STD" or "NONE" scaling (CoMFA_STD implies block standardization

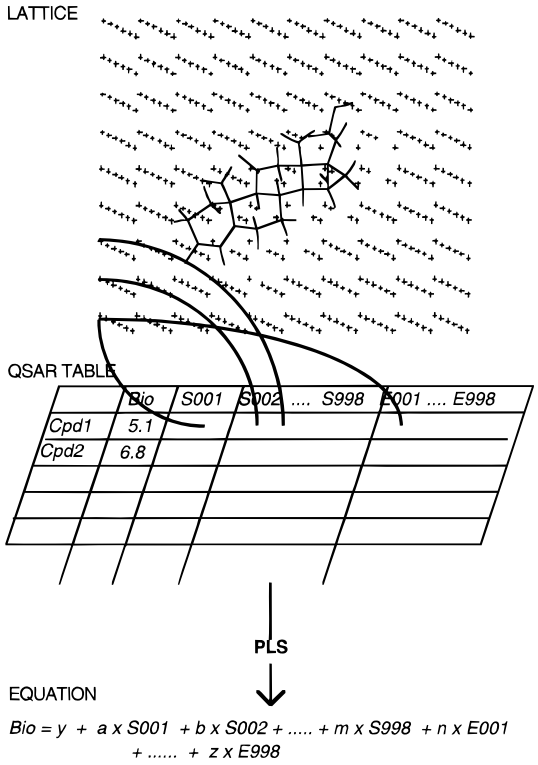



**Figure 2.** CoMFA process. "Steric field differences" are the root sum of squared differences between the corresponding cell values within a pair of rows in the QSAR table.

of each field but without rescaling of the individual "columns" corresponding to particular lattice points, which here produces the same clusters as no scaling). For clustering, the "distance" between any two compounds is calculated as the root sum of the squared differences in steric field values over all of the lattice intersections defined by the CoMFA "region". This distance calculation is completely equivalent to the calculation of a distance between points in geometric space; however, there are thousands of dimensions (lattice intersections) in a CoMFA field.

Both the steric field calculation and the field distance calculation can also be visualized in terms of the standard CoMFA schematic in Figure 2. The steroid Cpd1 has been positioned within the lattice, and the steric fields it exerts on hypothetical receptor atoms at the various lattice intersections are recorded within a row of a QSAR table, as shown by the arrows. In a standard CoMFA, PLS is used to generate a QSAR whose coefficients *a*,*b*,.. are largest wherever a change in the field value is most consistently related to changes in Bio. But in this study, referring again to the rows and columns in the QSAR table, the distance between any pair of compounds, for example Cpd1 and Cpd2, is calculated as described above. For example, the first term in the difference sum for this pair Cpd1,Cpd2 would be $(\text{Cpd1,S001} - \text{Cpd2,S001})^2$ and the second term $(\text{Cpd1,S002} - \text{Cpd2,S002})^2$.

The clustering level of 231 groups was chosen as the level which has the largest distance-to-the-next-level among all cluster levels from level 158 to 682. Put differently, the "stopping point" of 231 clusters, for which our main results are reported, represented the largest discontinuity in mean cluster separation encountered as the number of clusters decreased from 662 to 158.

## Results

The major result is that clustering of the steric fields for topomerically aligned conformers of 736 commercially available thiols[25] yields an esthetically very satisfactory division into 231 classes of bioisosteric structures. By visual inspection it was also tempting to propose for each of these bioisosteric classes a name

**Table 1.** Classifications for the 39 Largest of 231 Bioisosteric Groupings Found among 736 Commercially Available Thiols, by Clustering on the Steric Fields of Their Topomerically Generated Conformers as the Diversity Descriptor

| cluster ID | cluster size | struct root | struct substitution[a] | cluster ID | cluster size | struct root | struct substitution[a] |
|---|---|---|---|---|---|---|---|
| 1 | 26 | aryl | simple | 36 | 5 | 5-Het | 3,4-benzo |
| 33 | 5 | aryl | 2,3-benzo | 18 | 8 | 5-Het | 3-Ar |
| 7 | 14 | aryl | 2,6-NoH-3(4/5)-Me | 24 | 7 | 5-Het | 3-Me-5-H |
| 27[d] | 6 | aryl | 2,6-NoH-3-Ar | 35 | 5 | 5-Het | 3-Pe |
| 16 | 9 | aryl | 2-Et | 32[f] | 6 | 5-Het | 3-Pr |
| 37 | 5 | aryl | 3,5-Me | 34 | 5 | alkyl | simple |
| 10 | 16 | aryl | 3-Me | 3 | 18 | alkyl | (3:4) |
| 2 | 23 | aryl | 4-Me | 14 | 9 | alkyl | (3:4) (A1) |
| 19 | 8 | aryl | 6-NoH | 5 | 15 | alkyl | (5:7) |
| 12 | 10 | 5-Het | simple | 8 | 13 | alkyl | (8:11) |
| 22 | 7 | 5-Het | 2-(4-Et)Ar | 20 | 8 | alkyl | (10+) (B1) |
| 6 | 14 | 5-Het | 2-Ar | 15 | 9 | alkyl | (12+) |
| 39 | 5 | 5-Het | 2-Ar-3-(3-Ar)5-HetEt | 23 | 7 | alkyl | PheBu |
| 17 | 9 | 5-Het | 2-Ar-3-Ar | 25[b] | 6 | alkyl | PhePr |
| 28[e] | 6 | 5-Het | 2-Ar-3-Me | 11 | 10 | benzyl | simple |
| 30 | 6 | 5-Het | 2-Ar-3-PhePr | 9 | 13 | benzyl | 2/3-Me |
| 4 | 17 | 5-Het | 2-Me-3-(3-Bz)Ar | 21 | 7 | benzyl | 4-Et |
| 38 | 5 | 5-Het | 2-Me-3−5-Het | 26[c] | 6 | benzyl | 4-Me |
| 13 | 10 | 5-Het | 2-Me-3-Me | 31 | 6 | alkenyl | PEt3..Ar |
| 29 | 6 | 5-Het | 2-Me-3-PhePr | | | | |

[a] To generate these names, *all heteroatoms are first replaced by carbon* (to produce the simplest common topology) and a particular structure is chosen from among these topologies as the "most typical" of that cluster, if possible to contain the largest substructure that distinguishes that cluster from all others. Within the name of a substitution, numbers indicate positions when substitution is on a ring but chain length when substitution is on a chain (numbers separated by a colon indicate a range of chain lengths). Also, within a chain, letters indicate a position of substitution. (For example, (C2) describes a two atom branching from the third position of a chain, while 3-PhePr describes a phenyl propyl skeleton attached to the 3-position of a ring.) A dot notation (.) separates the three possible substituents on an alkenyl root, the substituent order being the same carbon as the −SH substituent and then the position *trans* to the −SH and finally *cis* to −SH. The above notwithstanding, *any* name enclosed completely in parentheses takes its usual structural meaning. Here are structural descriptions for each name abbreviation in the above table, mostly in SLN (SYBYL line notation), listed alphabetically. (SLN extends SMILES with the following concepts, among others. Hydrogens are explicit. Ring openings and closures begin with a number enclosed by [] and end with the matching number preceded by **@**. Other SLN symbols used in these SLN definitions are ∼, any bond; −, single bond (used here to provide a reference for [R]); :, aromatic bond; !, the SLN following (here in parentheses) is *not* allowed; [F], no additional atoms may be attached to the preceding atom; [!R], preceding bond may not be in a ring; [R], preceding bond must be in a ring.) 5-Het = C[1]:C:C:C:C:@1; alkenyl = C=C; alkyl = C∼[!R]C; aryl = Ar = Phe = Ph = C[1]:C:C:C:C:C@1; benzyl = Bz = HSC-[!R]C∼[R]C; Bu = C−[!R]C−[!R]C−[!R]C−[!R]C; cyclohexyl = Cyhx = C[1](−|=)C∼C∼C∼C∼C∼@1; cyclopentyl = C[1]∼(−|=)C∼C∼C∼C∼@1; Et = C−[!R]C; inden = C[1]:C(∼C∼X∼[2]):C(∼@2):C:C@1; iBu = C−[!R]C−[!R]C(−[!R]C)−[!R]C; iPe = C−[!R]C−[!R]C−[!R]C(−[!R]C)−[!R]C; Me = C; naphth = C[1]:C(∼C∼X∼[2]):C(∼@2):C:C:C@1; NoH = !(CH); o = denotes ring fusion, *e.g.*, benzo fuses a 6−membered aromatic ring; Pe = C−[!R]C−[!R]C−[!R]C−[!R]C−[!R]C; Pr = C−[!R]C−[!R]C−[!R]C; R# = alkyl chain of approximate length #; simple = !(C∼[!R]C); sPe = C(−[!R]C)−[!R]C−[!R]C−[!R]C−[!R]C; stilbenyl = C=[!R]C−[!R]C[1]:C:C:C:C:C@1; tBu = C(−[!R]C)(−[!R]C)−[!R]C. [b] See Figure 3 for the 2D and 3D structures from this cluster 25. [c] See Figure 4 for the 2D and 3D structures from this cluster 26. [d] See Figure 5 for the 2D and 3D structures from this cluster 27 [e] See Figure 6 for the 2D and 3D structures from this cluster 28. [f] See Figure 7 for the 2D and 3D structures from this cluster 31.

that is structurally both descriptive and distinctive. This naming scheme consists of two parts, the first part being the "root" substructure, *i.e.*, the fragment attached immediately to the −SH, and the second, the substitution pattern on that "root" substructure. The largest of these 231 classes is listed in Table 1, sorted by their proposed names. (The complete listing of all 231 clusters has been relegated to the Supporting Information.) A footnote to Table 1 provides additional decoding information needed for many of the class names. It will be noted that these names actually describe topologically equivalent hydrocarbons, *i.e.*, structures in which all monovalent atoms are discarded, and then hetero atoms are replaced by topologically equivalent carbons whose valences are then filled with the requisite number of hydrogens.

The smallest intercluster steric field difference at cluster level 231, the locally optimal stopping point according to strictly clustering criteria, is roughly 91.0 kcal/mol. This result proved useful for two reasons. First, and probably not coincidentally, 91.0 kcal/mol represents a steric field difference of roughly three grid points changing from occupied to unoccupied (30 kcal/mol having been used as the steric field cutoff), which in turn is about the average grid occupancy of a (nonattenuated) $-CH_2-$ group, the simplest synthetic building block. Second, and more likely fortuitously, the 91.0 kcal/mol value is very close to the average "neighborhood radius" of 85.0 kcal/mol found empirically from the neighborhood plots[19] for this descriptor.

The 231 clusters are rather homogenous in their population sizes, the largest two clusters, unsurprisingly, being "simple aryl" and "4-Me-aryl" with 28 and 23 members, respectively. ("Simple aryl", for example, includes fluorinated, chlorinated, and many aza derivatives of thiophenol.) There are 84 singleton clusters; in other words 11% of the 736 thiols are bioisosterically unique at cluster level 231.

To better illustrate the properties of these 231 clusters, Figures 3−7 show, for five representative clusters, the individual 2D structures and orthogonal views of their overlaid 3D topomerically aligned conformers. Specifically the clusters chosen are the most interesting from among those eight clusters which each contain six compounds, footnoted in Table 1. Both the 3D and 2D views are valuable. The clustering result itself of course depends only on the differences and similarities visible in the 3D overlay, but the 2D structures are much easier to examine individually. Within the 3D views, the anchoring −SH groups are at the left of the left panels.

Figure 3 (cluster 25) is the most homogeneous of the six from a 2D perspective. All "propyl" groups but one are actually $-CH_2CONH-$ groups. But the finding that the corresponding hydrocarbon, phenylpropyl, is also the

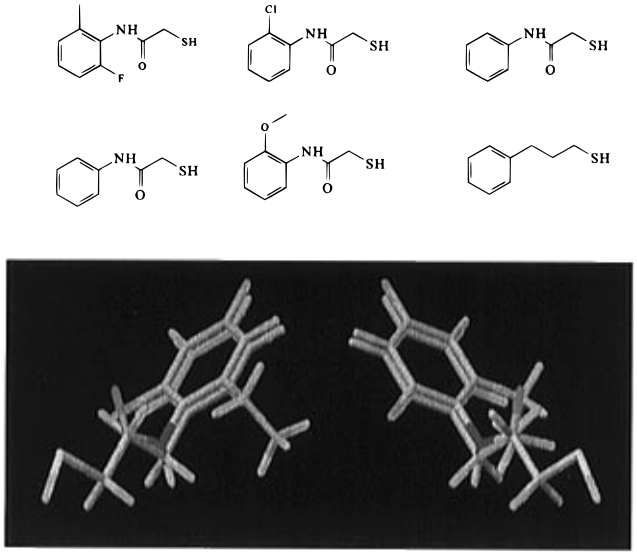


**Figure 3.** 2D structures and 3D overlays (orthogonal views) of the topomeric conformations for cluster 25, one of the 231 bioisosteric clusters found among 736 commercially available thiols.

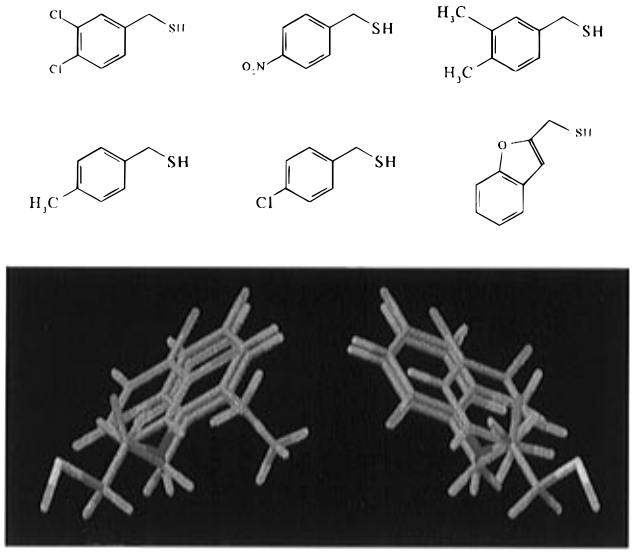


**Figure 4.** 2D structures and 3D overlays (orthogonal views) of the topomeric conformations for cluster 26, one of the 231 bioisosteric clusters found among 736 commercially available thiols.

sixth cluster member supports both the cluster name (within Table 1) and the overall naming procedure. The largest of the aryl substitutions that differentiate the five amides, $CH_3$ and $OCH_3$, might subdivide this cluster, were it not for both the attenuation of their field contributions because of four intervening rotatable bonds, reducing the effects of their atoms by almost one-half, and also the effectiveness of the topomeric alignment protocol in automatically overlaying the $CH_3$ and $OCH_3$ moieties, as shown in the 3D view of this cluster.

The cluster in Figure 4 (cluster 26) has many of the same features as that in Figure 3. All members of any one cluster are bioisosteres, of course, so benzofuran is being suggested as a bioisostere of a phenyl having small *meta* and/or *para* substituents. Also $NO_2$, $CH_3$, and Cl occupy much the same volume.

Figure 5 (cluster 27) differs from cluster 67 (not shown) only in the absence of any hydrogens *ortho* to the anchoring −SH moiety (all are 2-mercaptopyrimidines). This result exemplifies a general trend, that within any aryl ring, this steric field diversity measurement tends to separate *ortho* changes from those at *meta* and *para* positions, which instead tend to cluster together. This trend is a useful reflection of reality because *ortho* substitution always has the much stronger effect on the range of conformations accessible to a phenyl group.

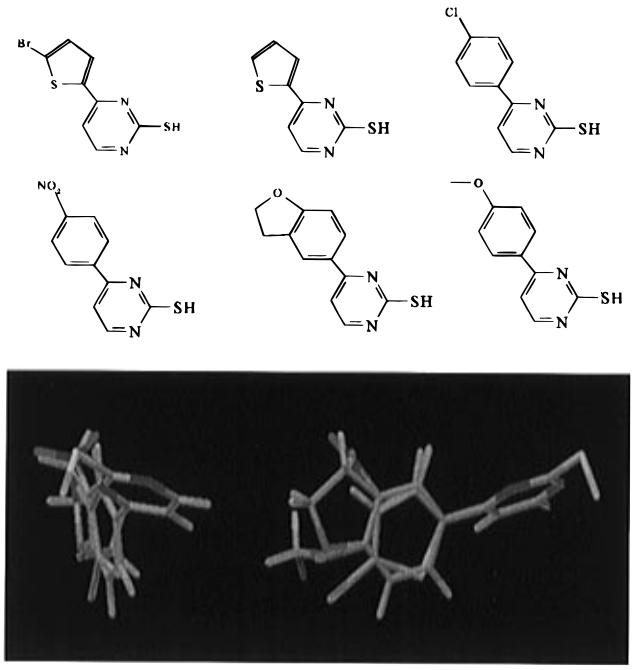
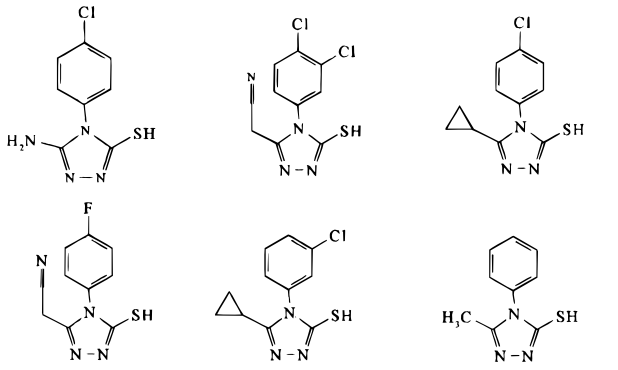


**Figure 5.** 2D structures and 3D overlays (orthogonal views) of the topomeric conformations for cluster 27, one of the 231 bioisosteric clusters found among 736 commercially available thiols.

Figure 6 (cluster 28) underlines another important generalization; 5-membered heterocycles are bioisosteric with phenyl rings *only* if their substituents are quite small. The reason, of course, is that a 3-substituent on a 5-hetero ring points about halfway between the directions of the 3- and 4-substituents on an aryl ring. When the substituents are small, their occupancy volumes almost completely overlap, but as the substituent atoms get farther from the root ring, their volumes spread apart and the steric field differences become large. Whenever sulfur is one of the atoms in a 5-membered heterocycle, its longer bond lengths create a similar though much smaller effect. For example, the structural class "3-aryl-5-Het" bifurcated into the two classes 18 (ring without sulfur) and 59 (ring contains sulfur).

Clusters 29 and 30 are not shown because their steric similarities are trivial outcomes of very high 2D structural similarity, while cluster 31 is omitted because its peculiar "alkenyl" root structure is not of known biological interest.

Figure 7 (cluster 32) is structurally the most heterogeneous of these five and also one of the more heterogeneous among all the 231 clusters. All of the root structures but one are thiadiazoles bearing a 3-chain three or four atoms long. But that sixth structure (lower right structure in Figure 7) has a partially saturated ring as its root (a root which would be named "cyclopentyl" in Table 1 nomenclature), and its 3-substituent is phenyl. One might expect that these two combined changes would force the sixth structure into a steric



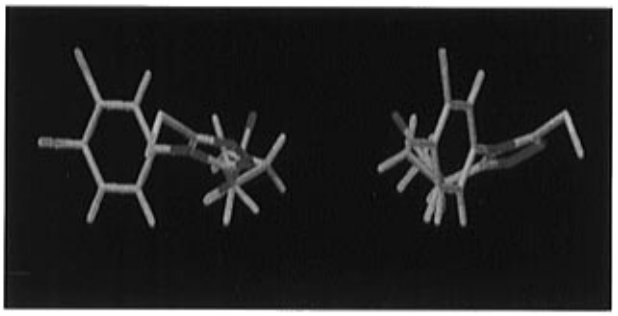
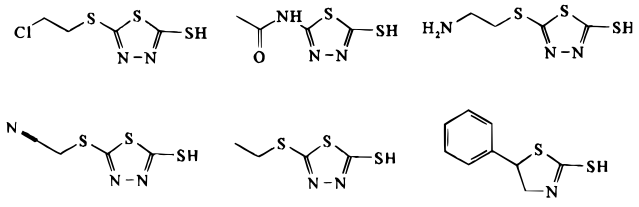


**Figure 6.** 2D structures and 3D overlays (orthogonal views) of the topomeric conformations for cluster 28, one of the 231 bioisosteric clusters found among 736 commercially available thiols.



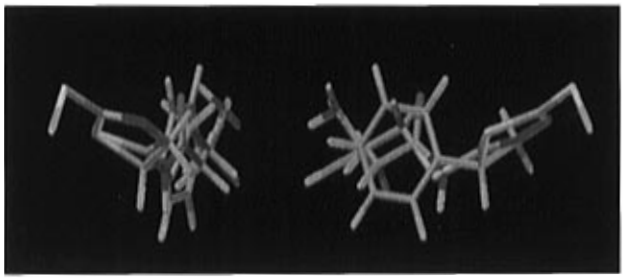


**Figure 7.** 2D structures and 3D overlays (orthogonal views) of the topomeric conformations for cluster 32, one of the 231 bioisosteric clusters found among 736 commercially available thiols.

diversity cluster all its own, but instead, as shown in the 3D overlay view, the phenyl ring fills much the same volume as does the "propyl" group of the other five compounds. In conformations other than the topomeric alignment used here, such as those that any particular receptor might induce, the sixth structure would assume a shape very different from the other five. Such an adventitious merging of two structure classes that must actually have quite different binding pocket preferences appears to be very rare. Individual inspection of all 231 clusters revealed less than a dozen such apparent artifacts.

To better understand the process of cluster formation, and to provide additional assurance that the reported

**Table 2.** Largest Differences between Consecutive Levels, in Hierarchical Clustering of Steric Fields of Topomerically Aligned Conformers of Commercially Available Thiols[a]

| differences | | | | | |
|---|---|---|---|---|---|
| >25.0, all levels | | >10.0, level > 7 | | >5.0, level > 42 | |
| level | diff | level | diff | level | diff |
| 1 | 66.9 | 8 | 15.6 | 50 | 5.3 |
| 2 | 55.6 | 13 | 19.5 | 54 | 6.9 |
| 5 | 27.0 | 20 | 10.1 | 88 | 5.4 |
| 7[a] | 25.5 | 35 | 10.4 | 97 | 7.4 |
| | | 42[a] | 11.2 | 109 | 5.3 |

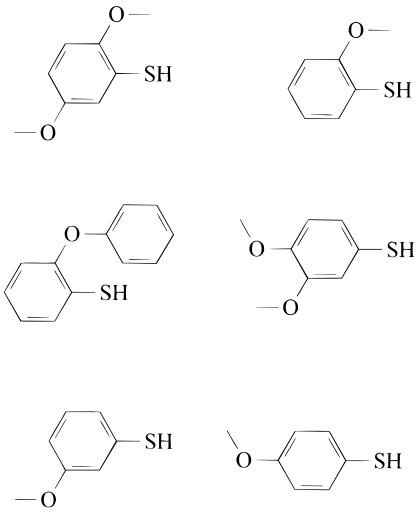
[a] Levels chosen for detailed visual examination, with results appearing in Table 2 of the Supporting Information.

231 level clustering was optimal, results obtained with fewer clusters were also examined. Levels of 7 and 42 clusters were chosen for examination in detail, both because these levels provided particularly large distances to the next level (see Table 2) and also to provide a rather even geometric ratio of spacing between the levels to be examined, approximately 6 ($6^3 = 216$, a bit less than 231).

The complete results from these classifications into 7 and 42 clusters, given in Table 2 of the Supporting Information, are summarized here. With seven clusters, 95.2% of the thiols fell into just three classes, distinguished much better by the general direction in which the side chain is aligned than by any necessarily very superficial chemical commonalties. With 42 clusters, although structurally-based naming begins to be of some value, cross-referencing individual clusters back to their various "children clusters" in Table 1 showed unacceptable structural heterogeneity.

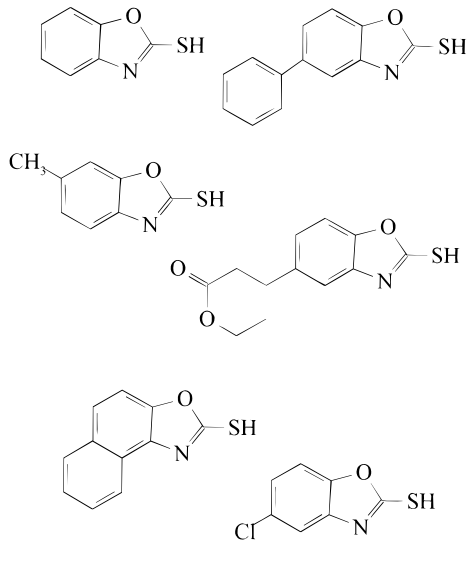
**Comparison with 2D "Fingerprint" Clustering.** Only one other diversity descriptor, Tanimoto coefficients of "2D fingerprints" of side chains, proved to be valid (exhibited a "neighborhood behavior") as often as did this "topomeric steric field" descriptor, in the neighborhood plot validation studies.[19] Thus for comparison we repeated the clustering of the 736 thiols into the same 231 classes using 2D fingerprints as the only descriptor. The overall distributions of class sizes are similar for the two descriptors. The largest cluster 2D fingerprint class includes 27 (long chain) thiols, and there are 80 singleton clusters. However there are important differences in how the individual compounds cluster.

We illustrate these differences by showing in Figures 8−12 the 2D structures for five of the seven 6-membered 2D fingerprint clusters. A useful way to contrast the different types of similarities reported by the two different descriptors is to ask for each of the clusters shown in Figures 8−12, "Within each of these five sets of six compounds reported as *similar* by their 2D fingerprints, how many fall into *different* clusters based on topomeric steric fields?" The largest divergence is in Figure 8 (cluster 24), with the six aryl ether derivatives being found in five different clusters when steric fields are the descriptor (in order, into clusters 80, 16, 152, 10, 10, and 2). Figures 9 and 11 (clusters 25 and 28) are almost as disparate, with the various benzisoxazoles and alkylated thiophenols each being found in four steric clusters (the mappings to steric clusters are 10, 161, 10, 169, 210, 10 and 88, 2, 48, 92, 92, 48, respectively). Figures 10 and 12 (clusters 26 and 29)



Cluster 24

**Figure 8.** 2D structures for cluster 24, one of the 231 clusters of structures similar by 2D fingerprints among 736 commercially available thiols.

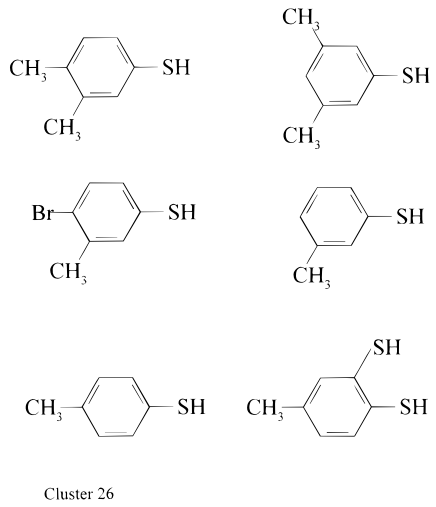


Cluster 25

**Figure 9.** 2D structures for cluster 25, one of the 231 clusters of structures similar by 2D fingerprints among 736 commercially available thiols.

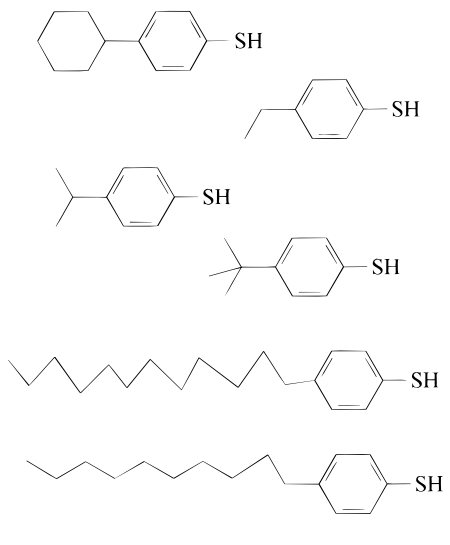
map into three steric clusters each, specifically 10, 37, 10, 10, 2, 10 and 131, 75, 75, 230, 75, 131.

To further emphasize the large steric field differences within clusters of structures found similar by 2D fingerprints, Figure 13 shows orthogonal 3D views of certain volume differences among the topomeric conformations of the compounds within 2D fingerprint clusters 24, 25, and 29. These volume differences compare each of the molecules that are not in the majority steric cluster with the union volume of the molecules in the majority cluster. For example, the cluster 24 panel of Figure 13 has four contours, showing the differences in volumes occupied by compounds **1**−**3** and **6** within Figure 8 from the volume occupied by either **4** or **5** (compounds **4** and **5** falling in the same cluster 10), in colors violet, red (obscured), blue, and green, respectively. Evidently the steric diversity within these three clusters is substantial, despite their obvious



Cluster 26

**Figure 10.** 2D structures for cluster 26, one of the 231 clusters of structures similar by 2D fingerprints among 736 commercially available thiols.



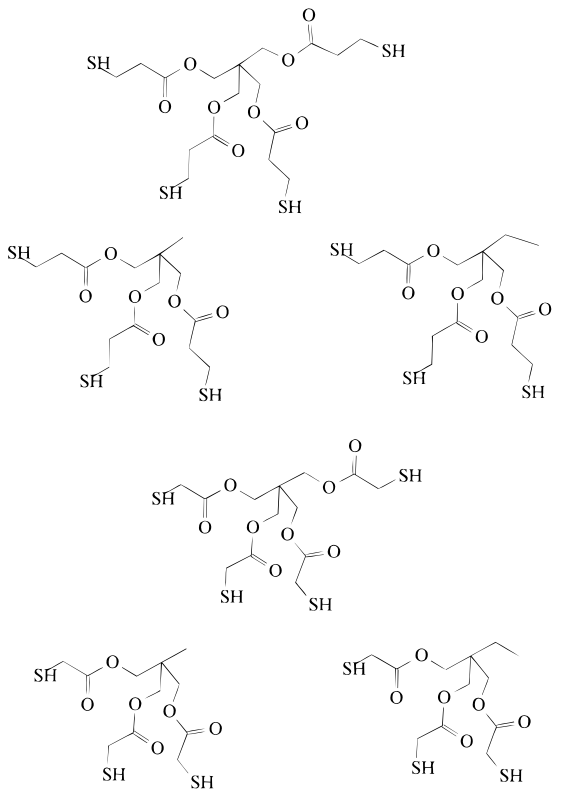
Cluster 28

**Figure 11.** 2D structures for cluster 28, one of the 231 clusters of structures similar by 2D fingerprints among 736 commercially available thiols.

similarities in 2D fragment composition. It seemed likely that when these side chains are constrained into a receptor cavity, the differences in steric field depicted by the volume differences would be decisive at least as often as the similarities in 2D fragment content, as was later confirmed by the neighborhood plot validation studies.

## Discussion

Several methodological choices were made in developing this topomeric steric field descriptor. In the end, the strongest justification for all these choices is empirical—the descriptors based on this topomeric alignment comprised two of the three most generally valid of 11 diversity descriptors, when evaluated using the neighborhood plots.[19] However three of the methodological choices made at the start of this study may merit further justification: first, the method of hierarchical clustering analysis with distances calculated by complete linkage, second, the use of steric fields only to



Cluster 29

**Figure 12.** 2D structures for cluster 29, one of the 231 clusters of structures similar by 2D fingerprints among 736 commercially available thiols.

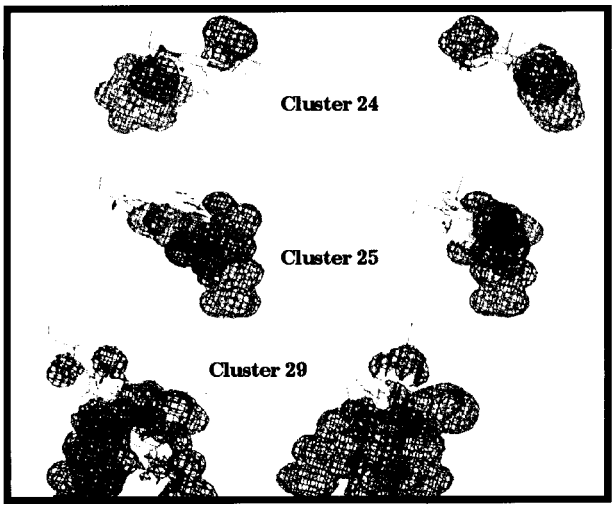



**Figure 13.** 3D overlays (orthogonal views) of volume differences among the topomeric conformations of compounds classified as similar on the basis of 2D fingerprints.

characterize side chains, and third, the representation of conformationally labile side chains by single conformers.

In general, cluster analysis seeks to partition objects (here, chemical structures) into classes or clusters based on the interobject distances. Usually the objects being clustered are for other reasons expected to form disjoint groups. For example, an arbitrary collection of biological organisms should be strongly clustered because several species are likely to be included. (Examples of an animal halfway between, say, a dog and a man, are rarely encountered—outside of legend!) However there is much less *a priori* rationale for why a set of chemical structures should necessarily form such "natural clusters" (which indeed is one of the reasons that we believe the incisiveness of these clustering results to be worth such emphasis).

The many clustering methods available vary mostly in how to select the next pair of clusters to merge (recalling that some of the "clusters" may actually be single objects) and how to define distances between clusters (which, usually being formed of multiple objects, permit many possible distance definitions). Our choice, hierarchical clustering, at each step selects next for merging the pair of clusters that are "closest", surely the most "natural" selection criterion! (On the other hand, hierarchical clustering then needs *all* intercluster distances, which becomes so costly as objects become numerous that other more approximate clustering methods become attractive.) Our other choice, "complete linkage", defines the distance between clusters A and B as the *maximum* of all distances a−b, where a is any object in A and b is any object in B. Overall, the intent of these choices is to maximize the similarities between all objects in the same cluster, at the cost of computational resources.

Contrasting approaches would be Jarvis−Patrick clustering or single-linkage hierarchical clustering. In the former approach,[26] clusters are formed based on the pairwise number of nearest neighboring clusters held in common. In the latter, the distance between clusters A and B is the *minimum* of all distances a−b, where a is any object in A and b is any object in B. Both these approaches tend to "chain", producing elongated clusters, for example, having objects p, q, and r, where p is close to q and q is close to r but p is quite distant from r.

Complete linkage hierarchical clustering tends to produce spherical clusters, whose positions remain stationary as individual objects are added and which themselves merge only reluctantly, in the last stages of the cluster analysis. This is what was observed. Table 2 shows that the smallest intercluster distances remain steeply increased at the end of the analysis. For these reasons we expect the "clusters" of Table 1 to be fairly robust in composition, rather independent of the particular compounds happening to be commercially available at any time, for example. This expectation is borne out by other work, to be reported elsewhere.

The original reason for using steric fields only was pragmatic. The resulting classification, that outlined in Table 1, seemed more compelling than when additional fields such as electrostatic or hydrogen bonding were included. Of course, a steric description alone must be incomplete, so why might it provide such better results? At least three answers can be suggested. The simplest is that steric interactions—classical bioisosterism—are certainly the best defined and probably the most important of the selective noncovalent interactions responsible for drug action. Another answer is that differences in electrostatic fields are not independent of differences in steric fields. For example, in order for a phenolic hydroxylic to exert some specific electrostatic effect on binding, there needs to be a phenyl group to sterically position the phenolic group. The final

answer begins with the observation that adding another field to the steric field will halve the steric contribution to the difference between one thiol "shape" and another, so that the very clean steric-only classification of the 736 compounds into 231 nameable clusters could be somewhat obscured. It is not that the additional field information is irrelevant, but that the 736 compounds are insufficient to provide as convincing a division into 231 classes when so many more measurements are allowed to contribute to the diversity.

With respect to conformer selection, there are two reasonable alternatives to our approach of considering a single conformer. One is to generate many "representative" conformers and compute a field from some average of these. Here, a difficulty is in believing that conformationally averaged fields will really be diagnostic for whether a group will discriminate between one receptor pocket and another. For example, twirling a naphthyl group and an anthracenyl group about their 1 and 9 bonds, respectively, must necessarily sweep out the same volume of space and so produce rather similar conformationally averaged fields, but when binding to a receptor, the naphthyl group will prefer an asymmetric pocket and the anthracenyl a symmetric pocket. The other alternative is to enumerate some *a priori* "representative set" of potential receptor pockets and classify side chains by which of these receptor pockets they fit inside. An example is the Chiron "polyominodocking" procedure.[12] Here the difficulties are in assembling the representative set of pockets, the limitation in discriminating diversity that any finite set of pockets must impose, and the handling of the constraining side-chain attachment bond.

These prospective difficulties persuaded us that the simpler approach of a topomerically aligned single conformer should be tried. The issue remains of whether the diversity patterns derived from such single conformers, however representative, usefully parallel the actual diversity in binding pocket specificities of the dynamic side chains. Nonetheless, by considering all of the displays exemplified in Figures 4−7, we have convinced ourselves that such a useful parallel indeed exists—that side chains in the same steric cluster would best fit into the same pockets and that side chains in different clusters could not be made to fit into all of the same pockets.

The topomeric alignment procedure described is very reliable in the absence of prochiral centers[23] and regioselectivity issues. None of the thiols required any manual realignment (nor did very many among the 20 validation series used in the neighborhood plot validation studies). On the other hand, some candidate validation series were discarded because of prospective stereocenters having unspecified chirality (including pyramidal nitrogen). In such ambiguous situations, Concord may build different stereoisomers for different series members, which when overlaid generate more steric diversity than actually exists. Nevertheless, the ability to generate reliably superimposable 3D models has value not only in combinatorial library design but also, for example, in promising automatic 3D-based QSAR analyses of the enormous volumes of structure−activity data which high throughput screening will generate.

An apparent limitation of this sterically dominated topomeric protocol is that the spatial positioning of some groups may seem less sensible whenever the resulting field descriptor is not primarily steric. For example, consider the hydrogen-bonding fields that would be generated for 2-thiolpyridines with bulky substituents in either the 4- or 6-position. The topomeric conformation will always deploy the bulkier substituent into the same region of space, thereby placing the hydrogen-bonding pyridyl moiety into either of two disjoint regions, which will then force the 4- and 6-isomers into two distinct hydrogen-bonding classes. Our opinion is that this differentiation into two classes is "correct", in that these 4- and 6-isomers will indeed have quite different receptor pocket affinities. In practice, however, the strongest two arguments for steric fields as useful diversity descriptors are the results of neighborhood plot validation[19] and the very strong agreement of the "bioisosteric" diversity classes produced by steric fields (in Table 1) with the intuition of experienced medicinal chemists and biochemists, as calibrated by decades of experimentation.

Why *does* this perhaps arbitrary-appearing set of procedures work so well? The strong agreement we have emphasized, between the results of this 3D modeling-based process and the perhaps largely 2D-calibrated perceptions of chemists, may suggest that our results have some hidden or trivial dependence on 2D similarity. However we doubt this. While our whole study shows that particular types of similarity in 2D structure will yield strong 3D similarities among their topomeric conformations, the occasional exceptions, such as the benzofuran among the 4-substituted phenyls in Figure 4 or particularly the thiazolidine interloper quite inappropriately clustered with the thiadiazoles of Figure 8, show that this classification fundamentally rests on 3D, not 2D, similarities.

Topomeric steric fields have three distinct components: the topomeric conformation, the use of CoMFA steric fields as shape descriptors, and the attenuation of atomic contributions by intervening rotatable bonds. It seems likely that other shape description techniques, such as the Gaussian functions introduced by Good and Richards,[27] could replace the CoMFA steric field with equal effectiveness. However alternatives to the other two components are not as easily visualized.

We believe that topomeric steric fields are complementary, rather than competitive, with 2D fingerprints. First, 2D fingerprints are readily applicable to almost any set of molecules (albeit with less general validity[19]), whereas steric fields as described here are applicable only to molecules having superimposable structural features, as produced by combinatorial synthesis. Second, because these two measurements appear orthogonal, as illustrated in Figure 13, then the two diversity measurements must be simultaneously useful. In our own design work on a combinatorial library for rapid lead discovery,[28] we currently require that each reactant be taken from a different steric cluster, following exactly the methodology described here, and also that every product has a Tanimoto similarity in 2D fingerprints of less than 0.85 with respect to every other product. We do expect that, with experience, steric fields and 2D fingerprints will usefully be augmented by other descriptors.

## Conclusion

Steric fields calculated for single "topomeric" conformations are strongly supported, by formal validation[19] as well as agreement with our collective experience of bioisosterism, to be very useful measurements of molecular diversity, especially suited for the design of combinatorial chemistry libraries sharing a common "core".[29]

**Acknowledgment.** We wish to thank the referees for several especially valuable comments.

**Supporting Information Available:** Version of Table 1 including all 231 clusters and similar information for the 7- and 42-cluster analyses, with all compound cluster memberships cross-referenced (9 pages). Ordering information is given on any current masthead page.

## References

(1) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. M. Measuring Diversity: Experimental Design of Combinatorial Libraries for Drug Discovery. *J. Med. Chem.* **1995**, *38*, 1431−1436.

(2) Madden, D.; Krchnak, V.; Lebl, M. Synthetic combinatorial libraries: Views on techniques and their applications. *Perspect. Drug Discovery Des.* **1995**, *2*, 269−285.

(3) Dean, P. M., Ed. *Molecular Similarity in Drug Design*; Chapman & Hall: London, U.K., 1995.

(4) Johnson, M. A., Maggiora, G. M., Eds. *Concepts and Applications of Molecular Similarity*; John Wiley: New York, 1990.

(5) Burger, A. *Progr. Drug Res./Arzneim.-Forsch.* **1991**, *37*, 287−371. Lipinski, C. A. *Annu. Rep. Med. Chem.* **1986**, *21*, 283−291.

(6) Cramer, R. D.; Redl, G.; Berkoff, C. E. Substructural Analysis. A Noveal Approach to the Problem of Drug Design. *J. Med. Chem.* **1974**, *17*, 533.

(7) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(8) Kubinyi, H., Ed. *3D QSAR in Drug Design. Theory, Methods, and Applications*; ESCOM: Leiden, Holland, 1993.

(9) Kim, K. H. Comparative molecular field analysis (CoMFA). In *Molecular Similarity in Drug Design*; Dean, P. M., Ed.; Chapman & Hall: London, 1995; pp 291−331.

(10) Carbo, R.; Leyda, L.; Arnau, M. An electron density measure of the similarity between two compounds. *Int. J. Quant. Chem.* **1980**, *17*, 1185−1189.

(11) Hopfinger, A. J. A QSAR investigation of DHFR inhibition by Baker triazines based upon molecular shape analysis. *J. Am. Chem. Soc.* **1980**, *102*, 7196−7206.

(12) Reviewed by Good, A. C. 3D molecular similarity indices and their application in QSAR studies. In *Molecular Similarity in Drug Design*; Dean, P. M., Ed.; Chapman & Hall: London, 1995; pp 24−56.

(13) Unger, S. H.; Hansch, C. *J. Med. Chem.* **1973**, *16*, 745.

(14) Van Waterbeemd, H. In *3D QSAR in Drug Design. Theory, Methods, and Application*; Kubinyi, H., Ed.; ESCOM: Leiden, Holland, 1993.

(15) Martin, E. J.; Blaney, J.; Siani, M.; Spellmeyer, D. Measuring diversity: Experimental design of combinatorial libraries for drug discovery. American Chemical Society Meeting, Anaheim, CA, 1995; COMP 32.

(16) Chapman, D.; Ross, M. J. Poster at the symposium Chemical and Biomolecular Diversity, San Diego, CA, Dec. 14−16, 1994; lecture at the symposium Exploiting Molecular Diversity: Small Molecule Libraries for Drug Discovery, La Jolla, CA, Jan. 23−25, 1995; conference summary available from Wendy Warr & Assoc., 6 Berwick Ct, Cheshire, U.K. CW4 7HZ.

(17) Jain, A. J.; Koile, K.; Chapman, D. Compass: Predicting Biological Activities from Molecular Surface Properties. Performance Comparisons on a Steroid Benchmark. *J. Med. Chem.* **1994**, *37*, 2315−2327. Jain, A.; *et al. J. Comput.-Aided Mol. Des.* **1994**, *8*. Jain, A. N.; Harris, N. L.; Park, J. Y. *J. Med. Chem.* **1995**.

(18) Marshall, G. R.; Barry, C. D.; Bosshard, H. E.; Dammkoehler, R. A.; Dunn, D. A. *ACS Symp. Ser.* **1979**, *112*, 205−226.

(19) Davies, K.; Briant, C. Combinatorial Chemistry Library Design using Pharmacophoric Diversity. Molecular Graphics Society Meeting, Leeds, U.K., April 1995.

(20) Teig, S. L. Diversity, Shmiversity: Is Your Library Any Good? CHI Meeting on Exploiting Molecular Diversity, San Diego, CA, Jan. 29, 1996.

(21) Brown, R. D.; Bures, M. G.; Martin, Y. C. Similarity and cluster analysis applied to molecular diversity. American Chemical Society Meeting, Anaheim, CA, 1995; COMP 3.

(22) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.

(23) A very recent extension of the topomeric alignment methodology currently allows reliable superposition of prochiral centers as well.

(24) Pearlman, R. S. 3D molecular structures: generation and use in 3D searching. In *3D QSAR in Drug Design. Theory, Methods, and Applications*; Kubinyi, H., Ed.; ESCOM: Leiden, Holland, 1993; pp 41−70. Program available from Tripos, Inc., St. Louis, MO.

(25) Available Chemicals Directory (ACD), 1995 version, available from MDL Inc., San Leandro, CA.

(26) Jarvis, R. A.; Patrick, E. A. Clustering using a similarity measure based on shared nearest neighbors. *IEEE Trans. Computers* **1973**, *C-22* (11), 1025−1034.

(27) Good, A. C.; Richards, W. G. Rapid evaluatin of shape similarity using Gaussian functions. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 112−116.

(28) Library of 100 000 compounds, currently being synthesized by PanLabs, Bothell, WA, and available from Tripos, Inc., St. Louis, MO. For more detail of the library design procedures themselves, see: Ferguson, A. M.; Patterson, D. E.; Garr, C. D.; Undereiner, T. L. Designing Chemical Libraries for Lead Discovery. *J. Biomol. Screen.* **1996**, accepted for publication.

(29) Patents pending.

JM960291F