# Measuring Molecular Similarity and Diversity: Total Pharmacophore Diversity

Gergely M. Makara*

*NeoGenesis Drug Discovery Inc., 840 Memorial Drive, Cambridge, Massachusetts 02139*

*Received January 24, 2001*

A novel method, total pharmacophore diversity (ToPD), based on known pharmacophore features for numerically defining molecular similarity or diversity is described. The method captures the 3D shape and functionality of molecules by the analysis of relevant intramolecular distances to generate a short and descriptive pharmacophoric fingerprint for each molecule. The ToPD fingerprints can then be used in diversity analysis, clustering, or database searching. Conformational sampling is carried out when needed by the means of molecular dynamics. Our results show that ToPD outperforms a traditional 2D fingerprint technique in all test cases.

## Introduction

Molecular diversity is a general concept that becomes quantitative when it is numerically defined to characterize the properties of small molecules. Classification of small molecules in drug discovery usually aims at improving hit rate in high-throughput screenings. Molecular recognition by macromolecules is largely mediated by shape and functional complementarity; thus, diversity methods that numerically describe molecules based on some representation of 3D shape and functionality are of particular value. Molecular diversity applications must handle a vast number of molecules; therefore, fast binary fingerprints are often used in comparisons of large databases. Two-dimensional methods, however, suffer from several major disadvantages. Lack of information on the actual shape and the location of the functional groups, poor recognition of isomers, and insensitivity to conformational issues can all render topological fingerprints fruitless for library design. Furthermore, combinatorial libraries are often composed of close scaffold analogues reacted with a series of building blocks along various projection vectors to scan receptor relevant diversity space. The products generated by such combinatorial syntheses can be representatives of unique 3D pharmacophores that are difficult if not impossible to differentiate by traditional 2D fingerprints.

Three-dimensional similarity techniques include three- and four-point pharmacophore methods,[1,2] surface-based methods,[3,4] and docking methods.[5] Pharmacophore methods generally identify pharmacophoric triangles or tetrahedrons and mostly ignore the total 3D shape of molecules.[1,2] Surface-based methods characterize molecular properties projected on some representation of the surface.[3,4] Molecules are then scored by docking surfaces of conformations of a molecule into the surfaces of conformations of another molecule. The pairwise docking process, due to its relative nature, needs to be carried out for all pairs because the method does not create a descriptive fingerprint for each molecule. Molecular hashkeys, computed by calculating the surface similarity of molecules compared to a basis set of molecules, eliminate this shortcoming and have been shown to correlate with molecular surface similarity.[6] The DOCK-based method measures molecular similarity as a function of predicted binding affinities against a reference panel of proteins.[5] The technique is a computational variant of Terrapin's affinity fingerprinting strategy.[7] Both affinity methods are limited to areas of diversity for which several proteins have been isolated and structural data exist. Alternative binding modes and potentially wide activity ranges add further complications to the analysis of affinity data. A novel concept, quantized surface complementarity diversity,[8] has recently been reported that compares molecules based on their ability to satisfy complementarity to protein surfaces. The model enumerates all theoretical combinations of quantized small molecule and protein surfaces at a low resolution.

In this paper, a new pharmacophore-based technique is introduced: total pharmacophore diversity (ToPD). The method works in distance space and uses distances between atoms of multiple conformations to describe the 3D shape and function of molecules. In our validation tests, ToPD-derived fingerprints outperform 2D Unity fingerprints in all cases.

## Total Pharmacophore Diversity

The concept of using pharmacophore recognition to describe the binding event of a small molecule to a macromolecule (for instance, a protein) has been employed for many years in the scientific literature. The ever-increasing number of protein−ligand crystal structures has both confirmed and expanded our understanding of molecular recognition. It is clear that the presence of important pharmacophore types in the right arrangement is very often required for a small molecule to bind to its target. In addition to pharmacophore points, however, surface-to-surface contact between ligand and target is established along the surface of the small molecule. Incompatibility in shape and function where close contacts exist may lead to significant loss of binding affinity even if the traditional pharmacophore point requirements are satisfied. This means that, for best results, the diversity method must consider the

* To whom correspondence should be addressed. E-mail: gregm@neogenesis.com. Phone: 617-588-5112. Fax: 617-868-1515

10.1021/jm010036h CCC: $20.00 
Published on Web 09/19/2001

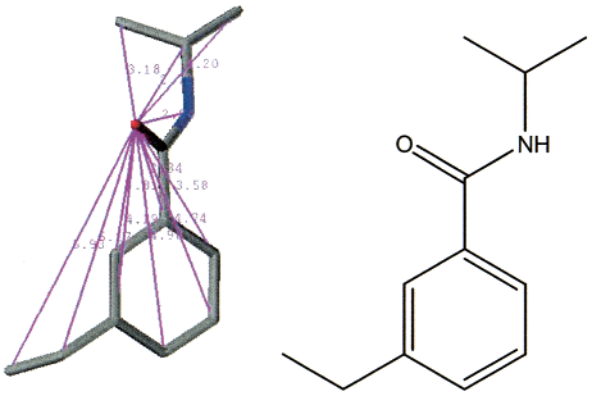


**Figure 1.** Distance map for an H-bond acceptor oxygen (left) for the structure shown on the right.

entire molecule as a whole. In addition to scoring matching pharmacophore patterns positively,[10−12] unmatched shape and functionality must be penalized to get meaningful predictions.

ToPD calculates pairwise distances between a defined set of atoms based on shape and pharmacophore type. Binned interatomic distances have previously been used as a descriptor type to define a geometric molecular similarity.[13] In this case, nonbonded intramolecular distances were partitioned into bins to derive a fingerprint, which was found to be similar to topological descriptors and showed no improved performance over its 2D counterpart.[13] The ToPD method extracts molecular shape *and property* information from the set of intramolecular distances in a unique approach.

Shape is captured by an ensemble of pairwise distances between all heavy atoms of the molecule. All other properties (currently hydrophobes, H-bond donors, H-bond acceptors, negatively charged, positively charged) are described by an ensemble of distances between the atom(s) that possesses the particular property and all heavy atoms of the molecule. The concept is demonstrated in Figure 1. In this fashion, the relative position of all pharmacophore features is mapped on the overall shape of the molecule. In other words, ToPD considers the location of the atoms within the molecule in relation to the overall shape of the molecule (which can be described by the positions of all heavy atoms). Thus, ToPD inherently penalizes for mismatches when molecules are compared as opposed to methods that score similarity solely based upon matching features. If the relative location of the same property for two different molecules is similar but the overall shapes are different, ToPD yields a low similarity value.

Distance values between two atoms can be attained based on a single conformation of the molecule or as an average of distances derived from several conformations of the molecule obtained by a conformational search method such as molecular dynamics. However, it should be noted that distance averaging across a diverse set of molecular conformations leads to meaningless values. A short molecular dynamics simulation in ToPD is applied to relax the starting conformation generated by 1D−3D conversion methods to a local energy minimum. Sampling of conformers for 30 ps is carried out only within this local energy well. Thorough sampling of the conformational space is not included in the current implementation of our method.

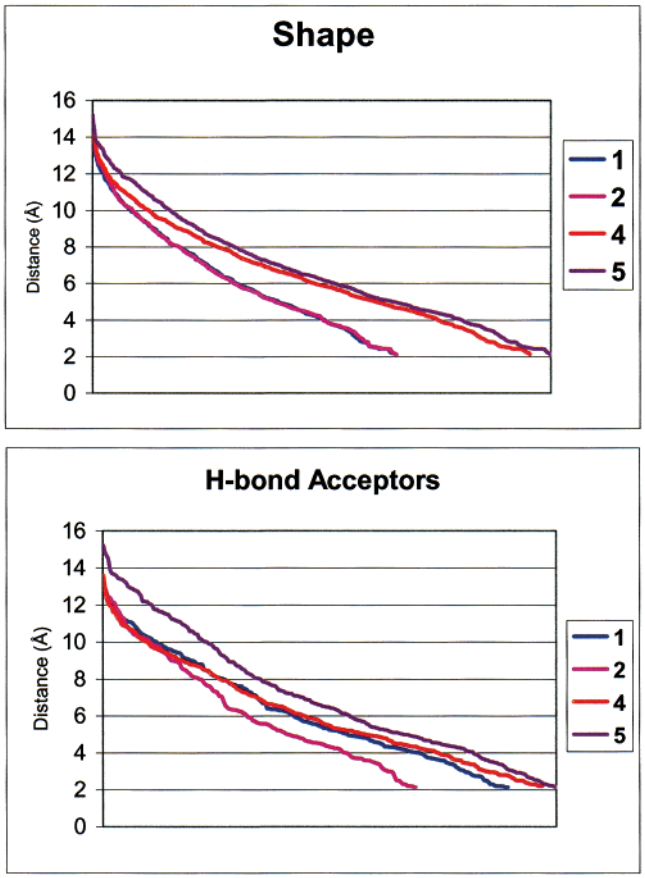


**Figure 2.** Examples of distance graphs for two pairs in Figure 3. Distance values are plotted in descending order.

Investigation of distance plots for test molecules unveiled that very short distances add only noise to the data. This derives from the fact that bond distances and three-atom angles are by nature highly redundant within organic compounds. All distances below 3 Å are, therefore, removed before analysis. Since ToPD works in distance space, the frame of reference for every molecule is internal and, therefore, no pairwise alignment is necessary when molecules are compared.

The set of distances that represent a particular property is sorted by increasing order of magnitude to yield a distance related plot for each molecule. Figure 2 depicts shape and H-bond acceptor functions obtained for pairs **1**−**2** and **4**−**5** (Figure 3). The shape plots overlap well for both pairs, but the H-bond acceptor curves reveal a variation for the **1**−**2** pair because **2** lacks H-bond accepting capability at the *o*-nitro region of **1**. It should be noted that shape plots for the two pairs are distinguishable despite considerable topological similarity between the two sets. The main difference is caused by the presence of a bulky *tert*-butyl group (**4**) and a phenyl ring in **5**.

When mathematically characterized, the atomic distance plots thus generated can express molecular recognition features. For each molecule, characterization values are extracted from the distance plots of each property type to yield a final string that we term a ToPD fingerprint. Characterization values may include slopes, intercepts, parameters of linear and nonlinear functions fitted on the distance plots, distance values, and number

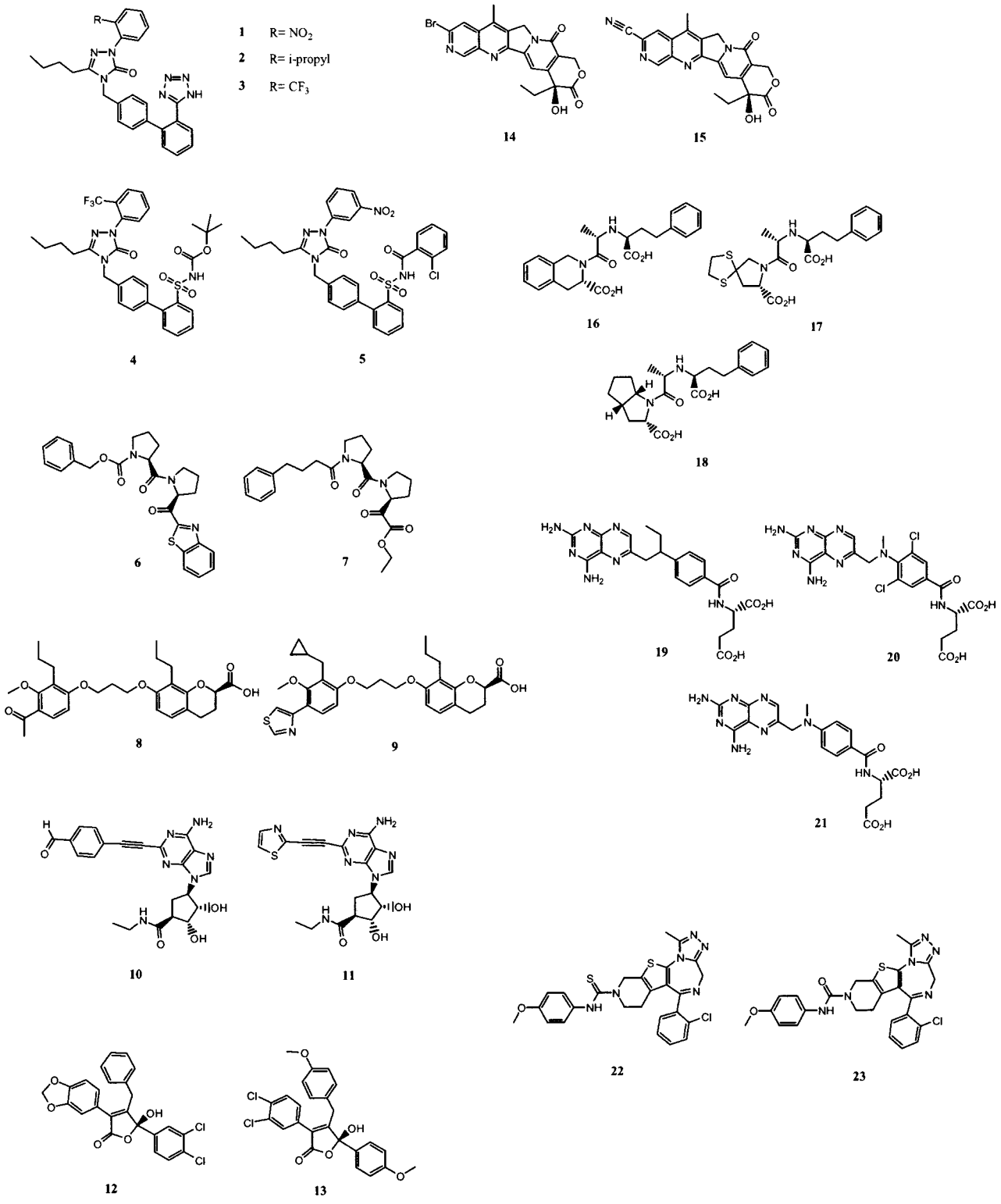


**Figure 3.** Molecules used in the close analogue series.

of distance values. The fingerprints can then be viewed as coordinates in a multidimensional space (where the number of dimension equals the number of fingerprint values in the string). Dissimilarity between molecules can be related to their weighted distance in this space: the farther apart the molecules, the more dissimilar they are. Different pharmacophore types may be weighted according to user-defined criteria depending on the application.

## Methods

Validation of ToPD has been performed on small data sets to enable careful evaluation of both positive and negative pairs, respectively. A small set of molecules was chosen that could be studied by means of biology, crystallography, and computational chemistry to yield the most reliable results possible. Once validated, the method can then be applied to problems of various sizes

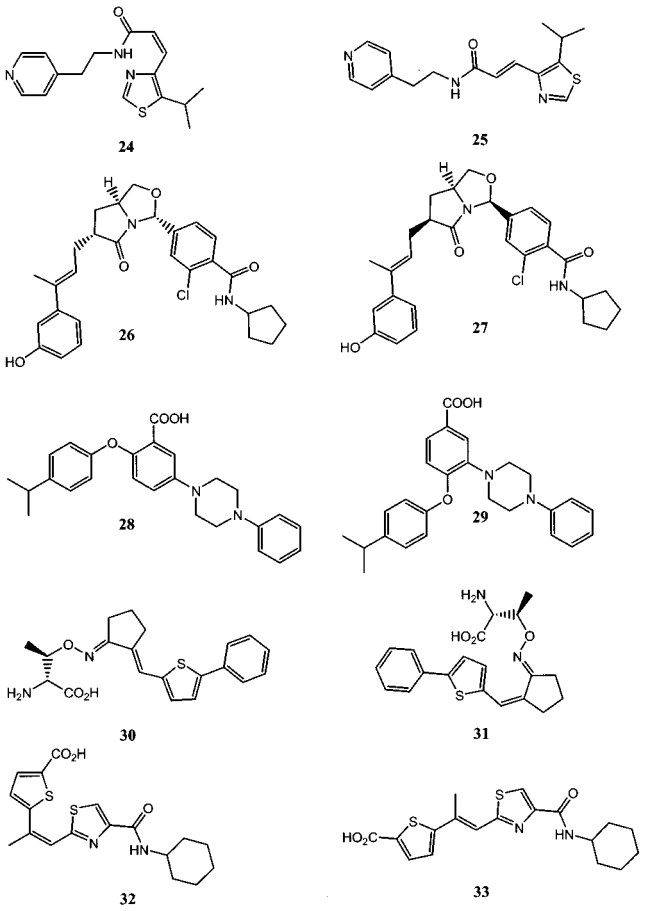


**Figure 4.** Structures of isomer pairs.

without a need for thorough assessment of every molecule pair.

Three test sets were designed to assess the effectiveness of ToPD compared to Unity fingerprints. Unity fingerprints have shown good neighborhood behavior and belong to a group of widely popular 2D methods along with Daylight fingerprints and Isis MOLSKEYS.[12] The first set is composed of series of very close analogues that are known to bind to the same target (Figure 3).[8] Two-dimensional methods are expected to perform extremely well in classification of molecules of such similar topology. A set of theoretical isomer pairs was also created to illustrate some of the deficiencies of binary fingerprints and to assess the ability of ToPD to resolve challenging pairs (Figure 4).

Most importantly a larger set of protein−ligand pairs deposited in the protein database was also evaluated. Fifty-six protein structures that passed analysis for similar binding modes[14] were selected (PDB filenames are listed in the Experimental Section). Selection of the crystal structures was initially based on a set previously published for the validation of the morphological similarity method.[4] This set was, however, found rich in very small molecules. Since our goal was to validate our method mainly on molecules similar to those likely to be present in our inventory, many pairs were removed from the reference set and replaced with PDB structures that contain established drug-like small molecules in the molecular range of 350−600. Such well-studied ligands include inhibitors of HIV protease, HIV reverse transcriptase, dihydrofolate reductase, and other enzymes. The size of ligand clusters that bind to the same site ranged from two to five. The presence of larger clusters is beneficial in order to increase the number of positives to a statistically higher level. Two or more ligands bound to the same active site form positives, while all other pairs are expected to be dissimilar (few exceptions to this general rule are discussed in the next section). Even with increased ligand clusters, the number of similar pairs is much lower than that of dissimilar ones. Conceptually, this ratio is desirable because it mimics large data sets of molecules (e.g., screening libraries) where the identity of a few positives is hidden among a vast number of negatives.

In the treatment of molecules, all ionizable moieties were converted to their predominating charged state under physiological conditions, and all validation tests were carried out using molecular dynamics as a conformational sampling tool. The set taken from the PDB was also used to measure the pharmacophore recognition power of ToPD and Unity fingerprints without the bias of conformer generation: use of CONCORD and molecular dynamics was compared to use of actual bound conformations. Separability (resolving power) of real positives (similar pairs) from real negatives (dissimilar pairs) provides a convenient means to demonstrate predictive power.[4] As an absolute and comparable measure of efficacy between different similarity methods, we examined the percentage loss of positives at the level of exclusion of 95% of all negatives.

## Results

Both ToPD and Unity fingerprints fair quite well in the close analogue series. ToPD fingerprints yield two nearly fully separated populations of positives and negatives, while Unity fingerprints give rise to a more diffuse distribution for similar pairs (Figures 5 and 6). Although the number of pairs in this study is rather small to make final conclusions, it is apparent from the false negative rate (Table 1) that ToPD performs at least as well if not better than a 2D fingerprint method even for molecules of analogous local topology. At the level of exclusion of 95% of all negatives, all positives are correctly identified by ToPD while 6% of similar pairs would be lost by the use of Unity fingerprints.

Isomer pairs were anticipated to provide the toughest test for ToPD because the method relies on differences in intramolecular distances. It is apparent that most short distances are identical for structural isomers; thus, successful recognition of pharmacophoric differences must be effected by sensitivity to variations in a relatively small number of medium- and long-range values. In addition, for structural isomers a significant portion of the overall shape of one of the isomers is identical to that of the other.

ToPD scoring for the structural isomer series outperformed Unity fingerprints (Table 2). This is likely a reflection of the fact that the 2D method does not contain bins for isomeric differences other than regioisomerism; diastereo and cis−trans isomers appear identical to Unity fingerprints. ToPD consistently succeeds in distinguishing regio, diastereo, and cis−trans isomers if the isomer switch results in appreciable changes in overall shape and functionality position. However, separability by ToPD does not seem to be correlated to

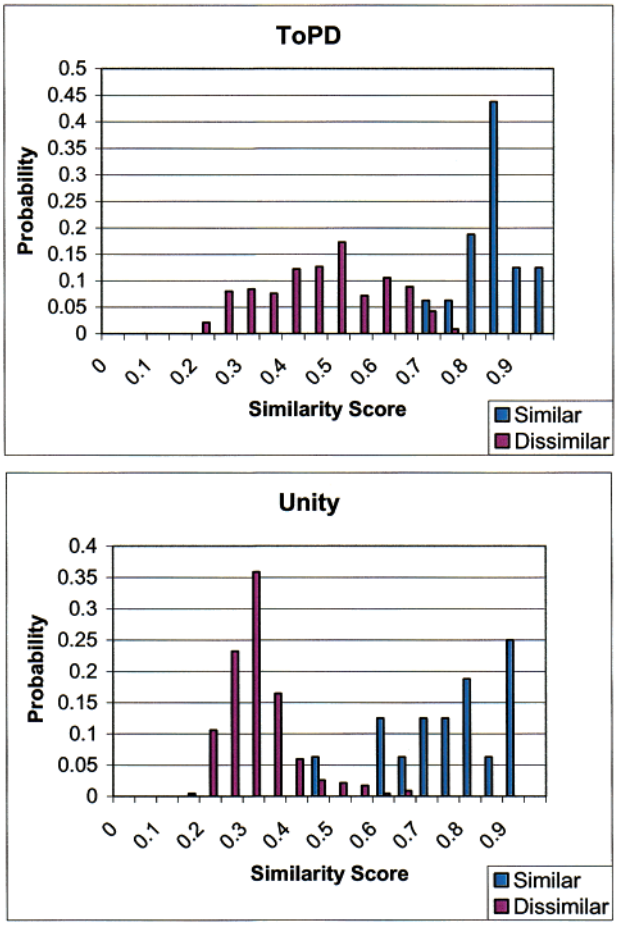


**Figure 5.** Normalized probability distribution of similar and dissimilar pairs obtained by ToPD and Unity fingerprints for the analogue series.

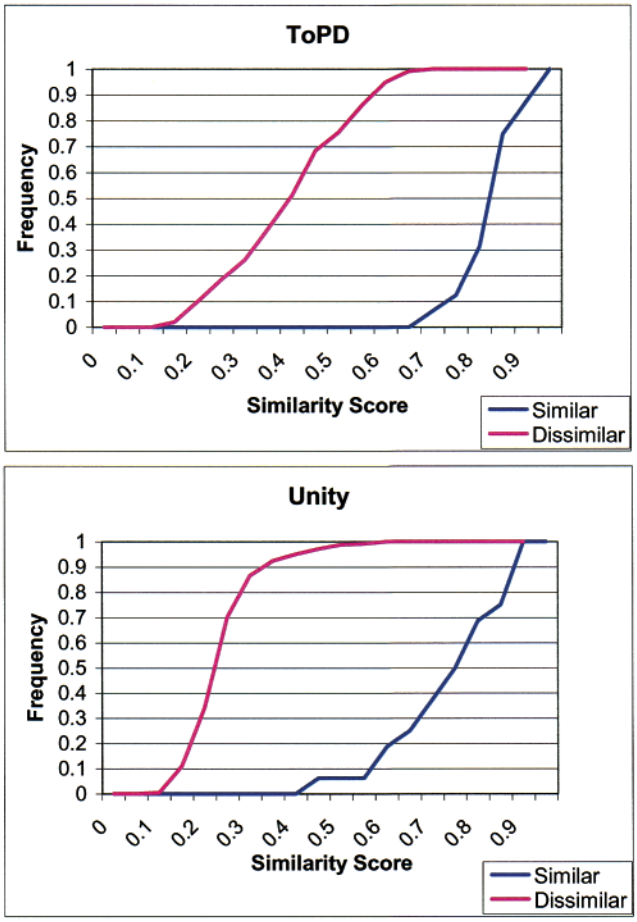


**Figure 6.** Cumulative distribution of similar and dissimilar pairs obtained by ToPD and Unity fingerprints for the analogue series.

**Table 1.** False Negative Rates at 0.05 Confidence Level (%)

| | close analogues | tailored PDB pairs[a] | tailored PDB pairs[b] | untailored PDB pairs[a] |
|---|---|---|---|---|
| ToPD | 0 | 3 | 14 | 17 |
| Unity | 6 | 31 | 31 | 41 |

[a] Without conformational search for ToPD. [b] With conformational search for ToPD.

**Table 2.** Similarity Scores and Surface Similarities for Isomer Pairs (%)

| | pair IDs | | | | |
|---|---|---|---|---|---|
| | **24−25** | **26−27** | **28−29** | **30−31** | **32−33** |
| ToPD | 0.73 | 0.73 | 0.78 | 0.45 | 0.40 |
| Unity | 1 | 1 | 0.84 | 1 | 1 |
| surface similarity[a] | 54 | 45 | 66 | 57 | 64 |

[a] Largest identical surface area as a percentage of total surface area.

surface similarity as measured by the largest identical surface area between two molecules (Table 2), suggesting that refinement of ToPD fingerprint parameterization may result in improvement in differentiation of structural isomers. Nevertheless, the current ToPD scores for this challenging series are lower than those of 85% of the real positives obtained for the close analogue series, and that appears to be an acceptable range considering that in most cases regio, diastereo, and cis−trans isomers do not bind to the same pocket.[15]

Protein−ligand pairs, the most relevant set for diversity validation,[14] were subjected to strict analyses in accordance with our validation set guidelines.[14] Initial visual inspection revealed the need for adjustment of 15 ligand structures to remove moieties that interact with protein binding cavities not involved in the binding of other ligands in the cluster.[14] This resulted in a "tailored set" (as opposed to the original ligand molecules that comprise the "untailored set"). For instance, COX-2 crystal structures include nonselective (**34**) and selective (**36**) inhibitors (Figure 7). Selective COX-2 ligands fill an additional binding pocket not utilized by the nonselective molecules. The tailored selective COX-2 ligand (**37**) interacts with the same residues as **35** does (Figure 8), and the pair of **35** and **37**, therefore, satisfies our diversity validation criteria for positives.[14]

First, ToPD fingerprints were calculated for binding conformations for each molecule in the tailored set without molecular dynamics. The results obtained therein are free of conformational bias due to 3D structure generation and conformational search and are a good measure of the separability power of the pharmacophore-based ToPD methodology itself. In a separate analysis, the ligand 3D structures of the tailored set were generated by CONCORD and subjected to molecular dynamics. This process mimics a typical ToPD procedure likely to be applied for diversity analysis of unknown sets. A rule-based structure generation method is unlikely to create 3D structures that perfectly match binding conformations; thus, performance of a pharma-

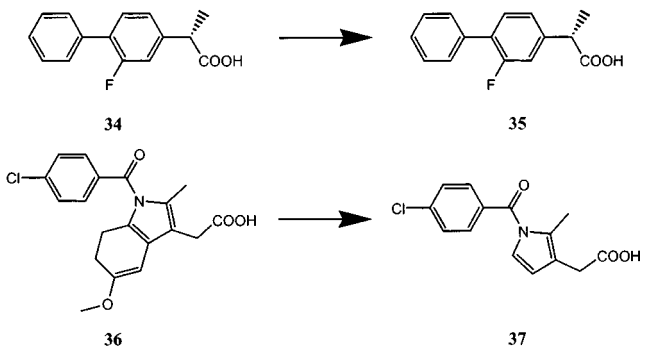


**Figure 7.** COX-2 ligands before and after tailoring.

cophore method is expected to drop while 2D data is independent of conformations. Finally, a set of fingerprints was produced out of the binding conformations of the untailored ligands to compare the ability of 2D binary and ToPD fingerprints to recognize false positives.

The tailored set of 57 ligands without conformational search resulted in 3% and 31% false negative rates for ToPD and Unity fingerprints, respectively (Table 1). The test set of 57 complexes consists of a few different enzymes that belong to the protease family. Some of these ligands showed similarity values over 0.7 between separate classes and visually appeared to be rather similar in shape and function. The selectivity of many of these ligands against other enzymes is not quantified; therefore, even these "false" negatives may be considered properly scored.

The above results indicate that the pharmacophore-based ToPD technique maintains a striking advantage in molecular recognition sensitivity when compared to 2D fingerprints. If the background for UNITY fingerprints is not known in a given study, the suggested cutoff of similarity of biological relevance for Unity fingerprints in the SYBYL manual is 0.85 or higher. However, *only 4% of the real positives meet that criterion in this study even though they are known to bind to the same areas of the same binding sites*. Even when the background for UNITY negatives is known, 31% of related pairs is lost if one wants to reject 95% of the dissimilar pairs (Table 1). The failure of Unity fingerprints is due to the diffuse distribution of real positives. The 2D binary method is unsuccessful in recognizing shape and functional similarity when connectivity relation does not exist. On the other hand, ToPD values are highly successful in separating related structures from unrelated ones (Figures 9 and 10).

Even when the 3D structures are generated by CONCORD and a molecular dynamics simulation is carried out, ToPD outperforms Unity fingerprints (14% vs 31% false negative rate) and gives 121% improvement in the false negative rate at significance level of 0.05 (Table 1). The background noise (negatives) remains unchanged (Figure 11) and the drop in the performance of ToPD is almost exclusively the result of decreased values in one active cluster. CONCORD generates very different starting orientations for molecular dynamics in the case of a cluster of five flexible peptidomimetic HIV protease inhibitors, resulting in a marked drop in the pairwise similarities. Since the favored state throughout molecular dynamics simulation for highly flexible molecules may greatly depend on the starting conformation generated by 1D−3D conversion, such flexible peptidomimetics will continue to pose difficulties for the current version of the ToPD method.

The final experiment carried out was a comparison of diversity scoring between the tailored and untailored validation sets. The 56 untailored ligands subjected to diversity comparison show a 14% and 10% hike in the false negative rates for ToPD and Unity fingerprints, respectively (Table 1). This makes sense since many of the presumed positives in this set are actually negatives if one examines observed cocrystal binding modes.[14] Three-dimensional fingerprints are very sensitive to these differences in relation to 2D fingerprints, suggesting that 3D methods are superior for diversity coverage of related target sites with varied local topology.

## Discussion

Total pharmacophore diversity as a diversity method encompasses several important advantages:

(1) ToPD generates a short shape and property related fingerprint file for every molecule, and the flexible format allows addition of new molecular recognition properties if needed. Description of features is not binary but continuous; thus, no errors can arise from a digital binning process. The fingerprint file describes the properties of a molecule itself, and calculation thereof needs to be carried out only once for every molecule, regardless of how many times that molecule is compared.

(2) ToPD considers an ensemble of all heavy (non-hydrogen) atoms to encompass the total shape and functionality as opposed to a few pharmacophore points considered by other methods. Molecules are compared in distance space without the need for alignment, and fingerprint files can be created with or without conformational search. A particularly useful application is when a single known binding conformation is used to give a ToPD fingerprint that can be compared to fingerprints of a series of molecules evaluated with molecular dynamics. The most similar structures identified by ToPD will not only have similar pharmacophore features but also preferred conformations close to the binding conformation of the known ligand.

(3) Similarity values in ToPD can be obtained for all included functionality individually. If a binding feature is suspected to be of particular relevance in a given study, its contribution to the overall similarity can be weighted accordingly or can be looked at separately. As a distance-based method, ToPD incorporates information on both overall molecular shape (long distances) and local topology (shorter distances) at the same time.

The disadvantages of the ToPD distance-based methodology are 2-fold. First, enantiomers are identical in distance space and currently are indistinguishable for ToPD. In case of a chiral hit, the identity of the more similar enantiomer must be established by other means. Second, since conversion of molecules to atomic distances is not necessarily a one-to-one mapping, it is possible that two very different molecules produce closely related distance plots that lead to overestimated similarity for the pair. In our experience this phenomenon is quite rare and has only a minor contribution to the false positive rate in ToPD.

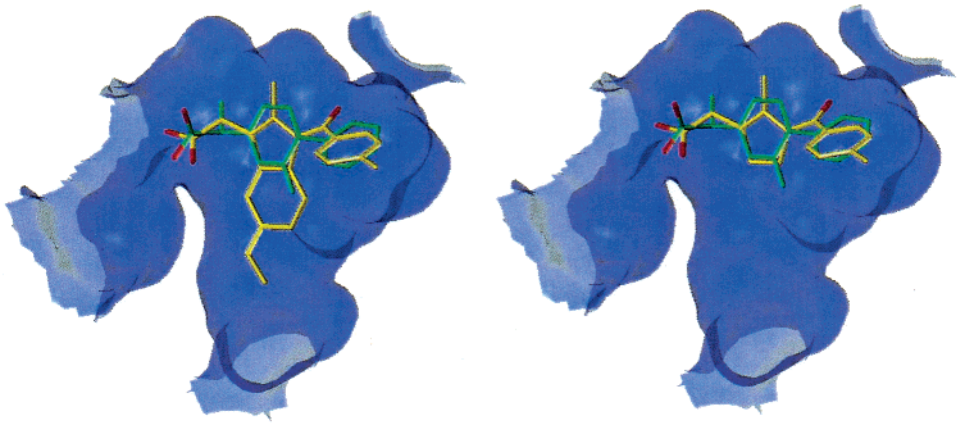

**Figure 8.** COX-2 ligands in the binding pocket before and after tailoring.

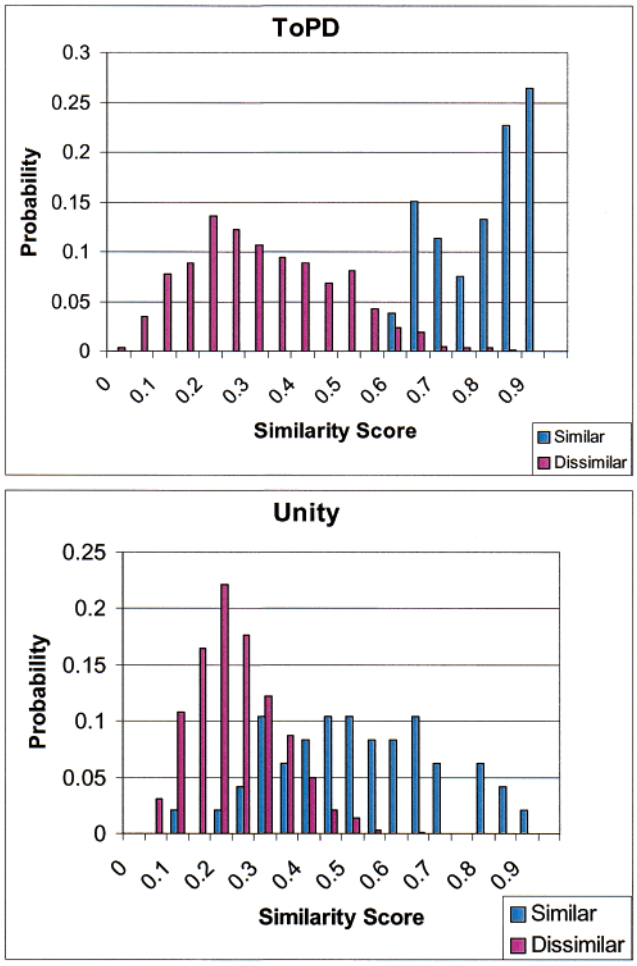


**Figure 9.** Normalized probability distribution of similar and dissimilar pairs obtained by ToPD and Unity fingerprints for the tailored PDB series.

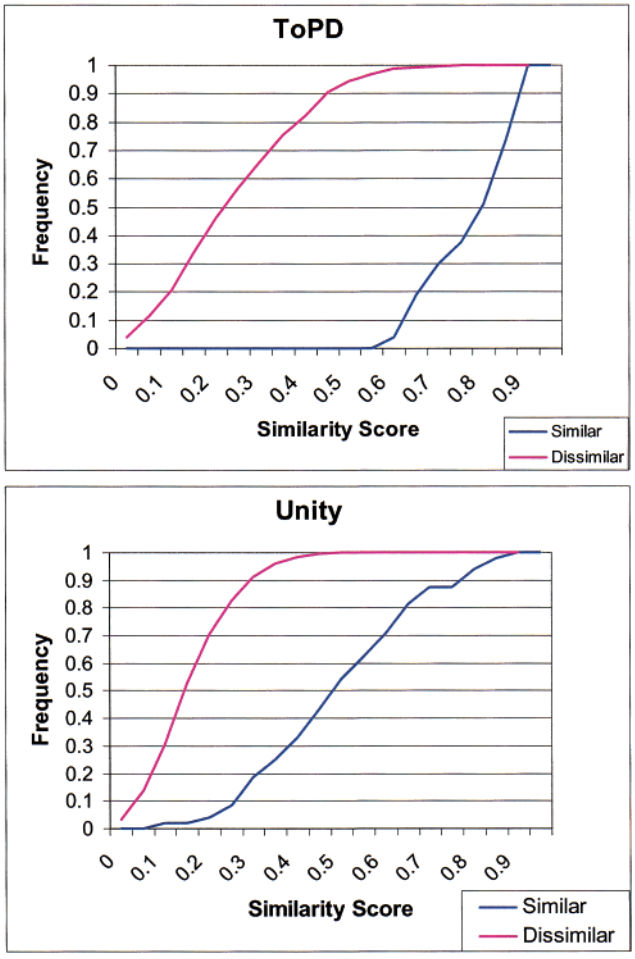


**Figure 10.** Cumulative distribution of similar and dissimilar pairs obtained by ToPD and Unity fingerprints for the tailored PDB series.

Validation tests reported in this paper reveal a substantial gain in recognition of important binding features by ToPD compared to a widely used 2D fingerprint method. ToPD classifies both positives and negatives correctly as opposed to Unity, which produces a diffuse linear plot for positives (Figures 5 and 9). A similar diffuse pattern was observed for Daylight fingerprints in a protein−ligand study.[4] Low similarity values are a result of the failure of the 2D method to correctly analyze analogues of unique topologies. While the 2D method works quite well for close analogues, it may be argued that these are of least interest, since for these pairs structural similarity is generally apparent without diversity analysis. While ToPD outperforms Unity fingerprints in both the tailored and untailored PDB series, comparison of performance of ToPD for the two PDB sets highlights the need for careful selection of diversity validation series.[14]

Introduction of conformational aspects leads to somewhat decreased performance of ToPD (Table 1). Although CONCORD tends to generate rather similar 3D

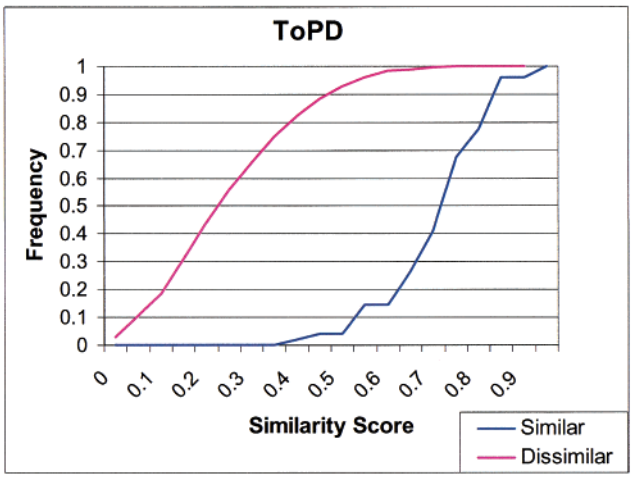


**Figure 11.** Cumulative distribution of similar and dissimilar pairs obtained by ToPD for the tailored PDB series with molecular dynamics.

structures for drug-like positive pairs, flexible molecules remain a challenge for ToPD. Molecules with many rotatable bonds can have a very large number of low-energy conformations and theoretically any reasonable minimum (within 10 kcal/mol of the calculated absolute minimum) on the potential energy surface can become a bioactive conformation. When the initial 3D structures generated by CONCORD for positive pairs are far from the actual binding conformations, and molecular dynamics simulations get trapped in local minima not related to these binding conformations *or to one another*, the similarity score calculated by ToPD for the pair is lower than expected. This shortcoming of the current implementation of our method can only be addressed by using comprehensive conformational search tools such as fast rule-base algorithms.[16] An improved methodology for fingerprint generation using all accessible conformations with no distance averaging is currently under development, and our results with the new algorithm will be reported at a later date.

ToPD fingerprints have routinely been utilized in our laboratory for screening library design at the core, building block, and product levels as well as for optimization libraries at the product level (small and large data sets, data to be published elsewhere). Cutoff values for positives for combinatorial libraries is usually set higher (0.85−0.9) than that for analysis of compound sets of high topological and functional diversity (0.7−0.75). Further improvements in the methodology are being implemented in the areas of isomer separation, conformational search, and particular molecular interactions observed in crystal structures.

## Conclusions

Findings reported herein suggest that 3D pharmacophore methods are superior to 2D binary fingerprints in molecular recognition tests. Total pharmacophore diversity is a distance-based method, which consistently and significantly outperforms Unity fingerprints for drug-like molecules. Highly flexible molecules pose difficulties for the current version of ToPD due to the lack of full conformational sampling. The method can be used for rapid evaluation of diversity in large screening libraries or generation of focused libraries as well as ligand-based virtual screening.

## Experimental Section

**Total Pharmacophore Diversity Method.** Structures were obtained either directly from PDB files or by 3D structure generation with CONCORD (Tripos, Inc.; St. Louis, MO). Molecules originated from PDB files were first manually checked and corrected for atom types in SYBYL 6.6 (Tripos, Inc). In-house SPL script was used to modify all atom types to ionized forms at physiological pH. Charges derived from the Gasteiger−Huckel method were applied before molecular dynamics simulations. In-house SPL script was used to submit all molecules to 50 ps molecular dynamics simulations with the Tripos force filed ($D$ = 78) at 300 K (20 ps equilibration followed by 300 conformations sampled over a 30 ps data collection period) and to analyze all trajectory files and store 300 conformations in a mol2 file for every molecule, respectively. An average 2 min/processor is necessary to complete a molecular dynamics run in SYBYL 6.6 on a dual processor Silicon Graphics R10,000 computer.

The output mol2 files (one bound conformation or 300 conformations from MD) were read by ToPD's fingerprint module to generate the fingerprint files that can be archived for later analysis. The list of atom types that covered the five pharmacophore types follows. Hydrophobic groups: $sp^3$ carbons; aromatic carbons; chlorines; bromines; iodines; $sp^3$ sulfurs; sp and $sp^2$ carbons if not attached to a heteroatom. H-bond donors: nitrogens or oxygens bonded to a hydrogen. H-bond acceptors: $sp^2$ and $sp^3$ oxygens; sp and $sp^2$ nitrogens; $sp^3$ nitrogens if not quaternary or positively charged; $sp^2$ and $sp^3$ sulfurs; fluorines. Positively charged: quaternary nitrogens; amidines; guanidines; carbocations. Negatively charged: oxygens of carboxylic acids (monocharged), sulfuric acids (monocharged), phosphonic acids (dicharged); nitrogen at position 2 in tetrazoles (monocharged); nitrogen in imides of C(=O)NS(=O)(=O) type. Interatomic distances were computed from Cartesian coordinates and sorted in descending order to give a distance function for each pharmacophore feature. The shape and property information can be extracted from the distance functions using various characterization values to give comparable results. In this study, molecular shape was described by seven descriptors: the slope and the intercept of the linear region (approximately the last 75% of the distance curves), the median distance value in the linear region, the slope and the intercept of the logarithm function of the nonlinear region (approximately the first 20% of the distance curves), the distance value at the end of the nonlinear region and the number of distances >3 Å. The hydrophobic term was described analogously to shape, but the number of distances was not used. Every other property was characterized by the slope and the intercept of regions of long (approximately the first 30% of the distance plot) and short (approximately the last 70% of the distance plot) distances, respectively. The values obtained therein were archived as ToPD fingerprints. Similarity values were obtained upon subjecting the fingerprints to similarity distance comparison by ToPD's similarity module. Similarities for individual properties (shape, hydrophobes, H-bond donors, H-bond acceptors, positively charged and negatively charged) were computed and stored separately and weighted equally to yield the final similarity number for all pairs. Fingerprints for shape and hydrophobicity were subjected to comparison and scaling as follows.

Similarity ($S$) was calculated for a fingerprint ($F$) value as

$$S = \frac{F2}{F1}$$

where $F1$ is the larger value.

This crude similarity value was scaled to give the final similarity measure for a fingerprint value

$$S = \frac{SA}{1 + (B/S)^8}$$

where $A$ and $B$ are scaling parameters and are dependent on the size of the larger molecule of the pair according to the following equations

$$A = 1 + \left[\frac{0.05}{1 + (700/D)^3}\right]$$

where $D$ is the number of distances in shape for the bigger molecule

$$B = 0.55 + \left[\frac{0.15}{1 + (700/D)^4}\right]$$

Thus, the fingerprint derived from the number of distances between heavy atoms (size) defined the weighting function and did not add to the dimensions of similarity space. This was required because our method evaluates differences as opposed to similarities. For instance, the relative dissimilarity caused by substituting a hydrogen with a phenyl group in a large molecule without size-scaling is considerably smaller than that in a very small molecule. The absolute contribution of a phenyl group to the binding energy, however, remains unchanged. Similarities for all other properties were obtained without weighting because the number of distances for features is *always* significantly less than that for shape and the number of these pharmacophore features is constrained to well-defined ranges.[17]

Similarity values computed separately for properties were averaged to give the final similarity measure between two molecules.

$$S_{\text{total}} = \frac{S_{\text{shape}} + S_{\text{hydrophob}} + S_{\text{HBA}} + S_{\text{HBD}} + S_{\text{poscharged}} + S_{\text{negcharged}}}{6}$$

If neither molecule of the pair contained functional groups belonging to a pharmacophore type, the shape similarity value was used for the feature during averaging in place of 1. This reflects the assumption that two molecules cannot be any more similar than their shape similarity because a match of molecular shapes is a prerequisite for binding to the same surface area of a given target.

ToPD fingerprints and similarity matrixes were calculated on a 600 MHz Pentium III PC (running Linux OS). Calculation of fingerprints for the current version of ToPD takes approximately 0.3−30 s/molecule depending on the number of conformers used (1−300) and the number of pharmacophore features present in the molecule. In our experience, a set of 50 sampled conformations gives rise to comparable results to that of 300 conformers. On average, the throughput of ToPD fingerprint module with 50 conformations for every molecule is about 15−20K molecules/processor/day. The nature of computation is highly parallel, and speed scales linearly with the number of processors added.

**Validation Tests.** For the close analogues and isomer pair series, all molecules were generated by CONCORD and subjected to molecular dynamics before ToPD analysis. Unity fingerprints were generated with Unity 4.1 (Tripos, Inc.; St. Louis, MO) and evaluated for Tanimoto scoring by an in-house perl script.

The PDB filenames used in the protein−ligand pair set follow (structures taken from an established data set[4] are shown in bold): 1hxb-1hsg-1hxw-**1hef-1phy**; **1cbs-1epb** (same ligand bound in different conformations); 1cht-1com; 1pge-3pgh-4cox; 1dbb-1dbl; 4dfr-1dr1-1jom-1hfq; **1did-1die-1xie**; **1dwd**-1a61-1a4w; **1fkg-1fki**; **1glp-1glq**; **1rob-1rsa**; **1srj-1stp**; 2aig-3aig-4aig; 2fke-1fkl; 1rt4-1rt5-1rt6-1rt7; 1hri-1ruc-1rud; 2qwk-1b9s; 3std-5std; 1fax-1xka; 1thl-1tlp; 1bx6-1ydt; 1bxo-1wea. The filenames for ligands modified in the tailored sets follow: 1hxb, 1hxw, 1hef, 4cox, 1dbl, 4dfr, 1fkg, 1fki, 1glq, 1rob, 1srj, 1rud, 1fax, 1bx6, 1bxo. For the tailored sets, 4dfr ligand was modified to give a pair with 1dr1 and was also left unaltered to pair up with 1jom and 1hfq. Hence, the total number of ligands in the tailored set was 57.

The surface of the binding pocket in Figure 8 was generated by MOLCAD with the default parameters in SYBYL 6.6. All molecule files and full similarity matrixes generated by ToPD in the paper can be obtained from the author upon request via e-mail (gregm@neogenesis.com).

**Acknowledgment.** The author thanks Tony Hopfinger (University of Illinois at Chicago) for the helpful discussions and Ed Wintner (NeoGenesis, Inc.) for his assessment of the manuscript. The assistance of Ciamac Cyrus Moallemi (NeoGenesis, Inc.) in scripting is greatly appreciated as well.

## References

(1) McGregor, M. J.; Muskal, S. M. Pharmacophore Fingerprinting. 2. Application to Primary Library Design. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 117−125.
(2) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-Point Pharmacophore Method for Molecular Similarity and Diversity Applications: Overview of the Method and Applications, Including a Novel Approach to the Design of Combinatorial Libraries Containing Privileged Substructures. *J. Med. Chem.* **1999**, *42*, 3251−3264.
(3) Mount, J.; Ruppert, J.; Welch, W.; Jain, A. N. Icepick: A Flexible Surface-Based System for Molecular Diversity. *J. Med. Chem.* **1999**, *42*, 60−66.
(4) Jain, A. N. Morphological Similarity: A 3D Molecular Similarity Method Correlated with Protein−Ligand Recognition. *J. Comput.-Aided Mol. Design.* **2000**, *14*, 199−213.
(5) Briem, H.; Kuntz, I. D. Molecular Similarity Based on DOCK-Generated Fingerprints. *J. Med. Chem.* **1996**, *39*, 3401−3408.
(6) Ghuloum, A. M.; Sage, C. R.; Jain, A. N. Molecular Hashkeys: A Novel Method for Molecular Characterization and Its Application for Predicting Important Pharmaceutical Properties of Molecules. *J. Med. Chem.* **1999**, *42*, 1739−1748.
(7) Dixon, S. L.; Villar, H. O. Bioactive Screening Library Selection via Affinity Fingerprinting. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1192−1203.
(8) Wintner, E. A.; Moallemi, C. C. Quantized Surface Complementarity Diversity (QSCD): A Model Based on Small Molecule-Target Complementarity. *J. Med. Chem.* **2000**, *43*, 1993−2006.
(9) Ajay; Murcko, M. A. Computational Methods to Predict Binding Free Energy in Ligand−Receptor Complexes. *J. Med. Chem.* **1995**, *38*, 4953−4967.
(10) Klebe, G.; Böhm, H.-J. Energetic and Entropic Factors Determining Binding Affinity in Protein−Ligand Complexes. *J. Recept. Signal Transduction Res.* **1997**, *17*, 459−473.
(11) Muegge, I.; Martin, Y. C. A General and Fast Scoring Function For Protein−Ligand Interactions: A Simplified Potential Approach. *J. Med. Chem.* **1999**, *42*, 791−804.
(12) 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.
(13) 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.
(14) For a detailed discussion on the disadvantages of inadequately verified data sets in diversity parametrization and validation, see: Makara, G. M.; Wintner, E. A. On Validating Similarity Metrics. To be submitted.
(15) For an exception, see Newcomer, M. E.; Pappas, R. S.; Ong, D. E. X-ray Crystallographic Identification of a Protein-Binding Site For Both All-Trans And 9-Cis-Retinoid Acid. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 9223−9227.
(16) For instance, see Omega from OpenEye Scientific Softwares.
(17) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental And Computational Approaches to Estimate Solubility And Permeability in Drug Discovery And Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

JM010036H