# Toward High Throughput 3D Virtual Screening Using Spherical Harmonic Surface Representations

Lazaros Mavridis,[†] Brian D. Hudson,[‡] and David W. Ritchie*[,†]

Department of Computing Science, King's College, University of Aberdeen, Aberdeen AB24 3UE, U.K., and
Centre for Molecular Design, Institute of Biomedical and Biomolecular Science, University of Portsmouth,
Portsmouth PO1 2DY, U.K.

Received April 28, 2007

Searching chemical databases for possible drug leads is often one of the main activities conducted during the early stages of a drug development project. This article shows that spherical harmonic molecular shape representations provide a powerful way to search and cluster small-molecule databases rapidly and accurately. Our clustering results show that chemically meaningful clusters may be obtained using only low order spherical harmonic expansions. Our database search results show that using low order spherical harmonic shape-based correlation techniques could provide a practical and efficient way to search very large 3D molecular databases, hence leading to a useful new approach for high throughput 3D virtual screening. The approach described is currently being extended to allow the rapid search and comparison of arbitrary combinations of molecular surface properties.

## INTRODUCTION

One of the main activities conducted during the early stages of a drug development project is to identify suitable drug leads. This often involves searching large chemical databases which may contain hundreds of thousands or even millions of compounds. Hence there is an ongoing need to develop more efficient and selective high throughput virtual screening (HTVS) algorithms.[1] Currently, the most commonly used techniques are based on binary bit-string representations of molecular properties and topologies such as Daylight, UNITY, and MACCS fingerprints.[2−4] However, by their nature, these representations have a tendency to find close chemical analogues to the given query, which may not be sufficiently novel to be worth pursuing in a drug development program (e.g., due to possible patent issues). On the other hand, a number of chemical databases exist which contain several hundred thousand three-dimensional (3D) molecular structures,[5] and this wealth of structural information could be better exploited by using improved methods of searching for similar 3D structures and substructures. For example, it is well-known that the pharmacological action of most drug molecules is governed by their interaction with their biological targets via ligand−receptor binding. Therefore, molecules that are globally similar in 3D space should share similar druglike properties. This suggests that using 3D comparison techniques should offer a better "scaffold-hopping" route to finding novel or unexpected drug leads.[6,7] However, comparing 3D molecular shapes and chemical properties is significantly more computationally expensive than comparing bit-strings. Developing methods for more efficient 3D comparison has recently become the subject of intense research.[8−18] Additionally, although typical corporate databases only contain a small fraction of molecules for which 3D structures have been solved crystallographically, it is common practice to store 3D structures generated from basic atom type and connectivity information using programs such as CORINA.[19] It is also now feasible to calculate molecular properties for entire data sets using semiempirical quantum mechanical techniques.[6] Hence it would be desirable to be able to search such data sets efficiently using shape-based and, ultimately, property-based queries as well.

There are many published techniques for superposing pairs of molecules.[1] Some simply maximize the volumetric overlap of the atoms in each molecule, but this can be expensive if calculated explicitly.[20] Other methods use reduced representations of molecular volumes,[21,22] surfaces,[23−25] or electrostatic properties[6,23,26] in order to reduce the amount of computation. For example, if Gaussian functions are used to represent atomic volumes, then the overlap between a pair of Gaussians can be calculated analytically, and molecular superpositions can be calculated in a matter of minutes.[10,27,28] As an alternative to using an explicit atom representation, Platt and Silverman used bipolar and quadrupolar components of molecular electrostatic fields to provide an efficient method of transforming similar molecules into a common frame.[26] Silverman extended this approach by placing a Dirac delta function at each atom center to calculate so-called steric multipoles which could be used as descriptors in a CoMFA analysis.[29] Steric multipoles are also used in the ROCS algorithm.[10] In ROCS, each molecule is prealigned with its inertial coordinate frame by diagonalizing a matrix of second-order steric multipoles calculated from atom-centered Gaussians. By minimizing atomic Gaussian overlaps for each of the four trial inertial starting orientations, ROCS is able to calculate up to 1000 molecular superpositions per second on a single processor.[17]

* Corresponding author e-mail: d.w.ritchie@abdn.ac.uk.
[†] University of Aberdeen.
[‡] University of Portsmouth.

10.1021/ci7001507 CCC: $37.00 © 2007 American Chemical Society
Published on Web 09/07/2007

In our Hex protein docking algorithm,[30] high order 3D spherical harmonic (SH) and Laguerre-Gaussian radial functions are used to give spherical polar Fourier (SPF) representations of macromolecular steric and electrostatic properties. Because such functions provide a complete orthonormal basis set, an infinitely high order expansion captures exactly the original properties. However, for practical purposes our SPF expansions are truncated at some limiting 3D polynomial order, typically $N = 25$. By exploiting the special rotational properties of the spherical harmonics,[31] Hex can evaluate millions of 3D overlap integrals per second in what is essentially a six-dimensional Fourier correlation search. However, because the translational part of a SPF correlation is relatively expensive, calculating an accurate molecular superposition using this approach takes several seconds on a contemporary personal computer, which is rather slow for HTVS purposes. Additionally, there are significant practical differences between the requirements and goals of docking and those of superposition calculations. For example, if one assumes that a pair of similar molecules may be adequately superposed and, perhaps, distinguished by colocating their centers of mass and by performing a pure rotational correlation of their shapes, then to a good first approximation each molecule may be represented very compactly using a two-dimensional (2D) SH surface envelope. Of course, using 2D surface envelopes assumes that the true molecular surface is starlike, or single-valued, with respect to radial rays projecting outward from the selected origin. Fortunately, however, this often holds to a very good approximation for small globular molecules. Even when this is not the case, it is nonetheless reasonable to suppose that similar molecules should give very similar radial projections and, therefore, that they should share very similar SH representations. This premise underlies our proposed SH surface based approach for 3D HTVS. It is worth noting that spherical harmonic surface representations are increasingly being applied to a broad range of object recognition and registration tasks in areas spanning, e.g., medical imaging, cryoelectron microscopy, and computer graphics.[32−35] Spherical harmonic surfaces have also been used recently to model protein−ligand shape complementarity.[11,36]

In this article, it is shown that the SH molecular surface envelopes contain sufficient information to superpose and classify small molecules rapidly and accurately, and the utility of using this approach to search molecular databases using simple scoring functions derived from this representation is investigated. The approach is demonstrated using a small database of 73 common drug molecules combined with some 1100 randomly selected decoy molecules. Our results show that using relatively low order SH correlations gives high recall and precision at very low computational cost. Calculations may be further accelerated with only a moderate reduction in precision by using rotation-invariant (RI) scoring functions and by comparing molecules in prealigned “canonical” orientations. The SH surface comparison approach is also compared with conventional physicochemical (PC) property clustering of both the drug data set and an additional data set of 46 odor molecules. In both cases, our results show that chemically meaningful clusters may be obtained using only low order SH expansions. The approach described has been implemented in a Java program called SpotLight.

## METHODS

**Shape Representation**. SH molecular surface envelopes are represented as radial distance expansions of the molecular surface with respect to a given origin, usually the center of mass

$$r(\theta,\phi) = \sum_{l=0}^{L} \sum_{m=-l}^{l} a_{lm} y_{lm}(\theta,\phi) \qquad (1)$$

where $(\theta,\phi)$ are the usual spherical coordinates, $y_{lm}(\theta,\phi)$ are real spherical harmonics, $L$ is the order or highest polynomial power of the expansion, and $a_{lm}$ are the expansion coefficients which are calculated as described previously.[31] The leading zero-order coefficient, $a_{00}$, essentially defines an average radius or “sea-level” for a molecule, and subsequent coefficients encode higher order local detail or “peaks and troughs” relative to this sea level. The leading coefficient normally has the largest magnitude, and subsequent coefficients distinguish molecules with similar sizes but different shapes.

Because the SH basis functions transform among themselves under rotation, it can be shown that SH molecular surfaces may be rotated by transforming only their expansion coefficients

$$a'_{lm} = \sum_{m'=-l}^{l} R^{(l)}_{mm'}(\alpha,\beta,\gamma) a_{lm'} \qquad (2)$$

where $(\alpha,\beta,\gamma)$ are $z$-$y$-$z$ Euler rotation angles and $R^{(l)}_{mm'}(\alpha,\beta,\gamma)$ are real rotation matrix elements which correspond to the Wigner rotation matrix elements for the complex harmonics.[37] Hence, having calculated a vector of expansion coefficients just once, a molecular surface may be reconstructed in an arbitrary orientation by substituting rotated expansion coefficients, $a'_{lm}$, into eq 1. In practice, however, reconstructing such surfaces is generally unnecessary because pairs of surfaces may be compared using the expansion coefficients directly. For example, if the rotation-dependent distance, $D_{\mathrm{ROT}}$, between a pair of SH envelopes, A and B, is calculated as

$$D_{\mathrm{ROT}} = \int (r_{\mathrm{A}}(\theta,\phi) - r'_{\mathrm{B}}(\theta,\phi))^2 \, \mathrm{d}\Omega \qquad (3)$$

then due to the orthogonality of the basis functions, and for a given limiting expansion order $L$, this reduces to

$$D_{\mathrm{ROT}} = \sum_{l=0}^{L} \sum_{m=-l}^{l} a_{lm}^2 + b'^{\,2}_{lm} - 2a_{lm}b'_{lm} \qquad (4)$$

Because the SH rotation matrices are also orthogonal, the first two squared terms are rotationally invariant. Only the final cross-term depends on the relative molecular orientations. This scoring function clearly has units of area, and a zero-order distance score corresponds to the difference between a pair of molecular sea level surface areas, for example.

**Rotationally Invariant Fingerprints**. Molecular superpositions can be calculated relatively quickly by minimizing eq 4. However, it is necessary to develop even faster comparison techniques in order to search very large 3D structural databases (e.g., $>10^6$ molecules) for HTVS. It is therefore natural to use the vector interpretation of SH coefficients to construct RI fingerprints (RIFs). Noting that expansion coefficients with the same value of $l$ transform among themselves under rotation, the RIF coefficients are defined as

$$A_l = \left(\sum_{m=-l}^{l} a_{lm}^2\right)^{1/2} \tag{5}$$

and

$$A_L = \left(\sum_{l=0}^{L} A_l^2\right)^{1/2} \tag{6}$$

By analogy to eq 4, the RIF distance score is written as

$$D_{\mathrm{RIF}} = A_L^2 + B_L^2 - 2\sum_{l=0}^{L} A_l B_{\mathrm{l}} \tag{7}$$

An even more efficient distance score based on RI magnitudes (RIMs) can be expressed as

$$D_{\mathrm{RIM}} = A_L^2 + B_L^2 - 2A_L B_L \tag{8}$$

**Data Set Selection**. In order to test our approach, three different groups of molecules were assembled, here called Drug, Odor, and Random. The Drug data set was compiled from the intersection of the RxList list of top 200 prescription drugs of 2002[38] and the Chembank bioactive database[39] to give a list of 73 compounds covering several of the main pharmacological drug classes, with several representative structures from each class. The molecules in this data set have molecular weights (MWs) ranging from 129 to 557 Da and are listed in Table 1. The Random data set was assembled from the NCI database[40] by randomly selecting 1108 molecules with MWs within the above range but excluding molecules containing any metal atoms in order to provide a reasonably comparable set of decoys. Figure 1 shows the MW distributions of the Drug and Random data sets. The Odor data set comes from Takane and Mitchell.[41] The 3D structures of all compounds were calculated using CORINA.[19]

**Database Search Analysis**. In principle, very high order vectors of SH expansion coefficients could be used to identify uniquely each molecule in a database. However, because the first few expansion coefficients are the most significant, it follows that using truncated expansions should give a good compromise between time efficiency and precision/recall behavior, thus facilitating searches over potentially very large numbers of molecules within a database. In other words, it seems reasonable to expect that the best matches for a given query molecule should be retrieved rapidly using only a relatively short SH coefficient query vector. Additionally, performing a HTVS search involves scanning a database for molecules which are similar but not identical to the given query. Thus a further advantage of the SH representation is that the expansion order, $L$, provides a convenient way to control the level of resolution at which different molecules may be compared.

In order to assess the utility of the above SH scoring functions in a simulated HTVS search, a small subset (2%) of the Drug plus Random database were defined as "hits" with respect to an arbitrarily selected query molecule, lorazepam (MW=321 Da). The hits were defined by their similarity to lorazepam when calculated using near-infinite $L = 15$ SH expansions with 5° search steps, which is treated as our "gold standard" similarity measure. A hit threshold value of 2% was chosen in order to provide a reasonably large and diverse set of molecules with which to conduct the database tests. Receiver-Operator-Characteristic (ROC) curves[42] were then used to analyze the precision/recall behavior of the scoring functions with respect to these predefined hits using a range of expansion orders and search step sizes. In the ROC terminology, our gold standard hits are treated as "positives" while the remaining 98% constitute "negatives". In the following database query tests, a true positive (TP) was assigned when the calculated hit list (i.e., the top 2%) includes one of the original positives, and a false positive (FP) was assigned when the hit list includes one of the original negatives. A true negative (TN) occurs when an original negative appears outside the calculated hit list. Similarly, a false negative (FN) occurs when an original positive falls beyond the calculated hit list. ROC curves are plotted here with the true positive rate (TPR) on the $y$-axis against the false positive rate (FPR) on the $x$-axis, where TPR and FPR are calculated as

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{9}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{10}$$

## RESULTS

**Determining Optimal Search Parameters.** The two most significant factors that control the speed and accuracy of SH comparisons are the Euler angle search step size and the order of the expansion. To determine a good step size, pairwise distances of lorazepam with the remaining molecules in the Drug database were calculated. Figure 2 shows plots of SH distance as a function of the rotational step, $S$, using a moderately high expansion order of $L = 6$. This figure shows that the distance score is almost constant for search step sizes up to 15°. Hence this value defines the maximum step size which should be used for reliable rotational comparisons.

In order to find the lowest value of $L$ which still gives good superpositions for arbitrary shape comparisons, one large and one small molecule from the Drug data set were selected to act as representative query molecules against the remaining set. Those two molecules were then compared against the remainder of the Drug data set by calculating pairwise distances for a range of expansion orders. The curves in Figure 3 show that the distances between the query molecules and each of the remaining database molecules flatten to a near-constant value when $L \geq 5$. Because none of the curves cross each other when $L \geq 5$, it can be concluded that $L = 6$ is sufficient to rank reliably all of the database molecules in order of distance for both large and small query molecules. Figure 3 shows that often $L = 3$ is sufficient to identify a small number of very good matches, but that $L = 6$ should be used to cover all cases. Figure 3 also shows the corresponding curves obtained for the RIF and RIM scoring functions. In these curves, the RI scores

**Table 1.** Pharmacological Classification of the 73 Molecules of the Drug Data Set[a]

| name | class | World Drug Index keywords |
|---|---|---|
| minocycline | AB 1 | antibiotics |
| doxycycline | AB 1 | antibiotics |
| tetracycline | AB 1 | antibiotics |
| cefprozil | AB 1 | antibiotics |
| clindamycin | AB 1 | antibiotics |
| ibuprofen | AI 1 | analgesics; antiinflammatories; antipyretics |
| aspirin | AI 1 | analgesics; antiinflammatories; antipyretics; anticoagulants |
| diclofenac | AI 1 | analgesics; antiinflammatories; prostaglandin-antagonists |
| naproxen | AI 1 | analgesics; antiinflammatories; prostaglandin-antagonists; antipyretics |
| codeine | AI 1 | analgesics; antitussives; narcotics |
| carisoprodol | AI 1 | analgesics; relaxants |
| loratadine | AI 2 | antihistamines-H1 |
| cetirizine | AI 2 | antihistamines-H1 |
| promethazine | AI 2 | antihistamines-H1; sedatives |
| triamcinolone | AI 3 | corticosteroids |
| methylprednisolone | AI 3 | corticosteroids |
| budesonide | AI 3 | corticosteroids |
| prednisone | AI 3 | corticosteroids |
| clonazepam | CN 1 | anticonvulsants |
| gabapentin | CN 1 | anticonvulsants |
| phenytoin | CN 1 | anticonvulsants |
| topiramate | CN 1 | anticonvulsants |
| sertraline | CN 2 | antidepressants; psychostimulants |
| fluoxetine | CN 2 | antidepressants; psychostimulants |
| nortriptyline | CN 2 | antidepressants; psychostimulants |
| amitriptyline | CN 2 | antidepressants; psychostimulants |
| paroxetine | CN 2 | antidepressants; psychostimulants |
| citalopram | CN 2 | antidepressants; psychostimulants |
| bupropion | CN 2 | antidepressants; psychostimulants |
| olanzapine | CN 3 | psychosedatives; dopamine-antagonists; neuroleptics |
| risperidone | CN 3 | psychosedatives; neuroleptics; antiserotonins; dopamine-antagonists |
| lorazepam | CN 3 | psychosedatives; tranquilizers |
| buspirone | CN 3 | psychosedatives; tranquilizers |
| diazepam | CN 3 | psychosedatives; tranquilizers |
| temazepam | CN 3 | psychosedatives; tranquilizers; anticonvulsants |
| trazodone | CN 3 | psychosedatives; tranquilizers; psychostimulants; antidepressants |
| cyclobenzaprine | CN 3 | psychosedatives; tranquilizers; relaxants |
| zolpidem | CN 3 | psychosedatives; tranquilizers; sedatives |
| fenofibrate | CV 1 | antiarteriosclerotics |
| gemfibrozil | CV 1 | antiarteriosclerotics |
| simvastatin | CV 1 | antiarteriosclerotics; HMG-COA-reductase-inhibitors |
| pravastatin | CV 1 | antiarteriosclerotics; HMG-COA-reductase-inhibitors |
| warfarin | CV 2 | anticoagulants |
| nifedipine | CV 3 | cardiants; calcium-antagonists |
| diltiazem | CV 3 | cardiants; calcium-antagonists |
| verapamil | CV 3 | cardiants; calcium-antagonists; protein-kinase-C-inhibitors |
| triamterene | CV 4 | diuretics |
| spironolactone | CV 4 | diuretics; aldosterone-antagonists |
| hydrochlorothiazide | CV 4 | diuretics; carbonic-anhydrase-inhibitors; hypotensives |
| furosemide | CV 4 | diuretics; protein-kinase-C-inhibitors |
| valsartan | CV 5 | hypotensives |
| terazosin | CV 5 | hypotensives |
| captopril | CV 5 | hypotensives; angiotensin-antagonists |
| fosinopril | CV 5 | hypotensives; angiotensin-antagonists |
| doxazosin | CV 5 | hypotensives; sympatholytics-alpha |
| bisoprolol | CV 5 | hypotensives; sympatholytics-beta; antiarrhythmics |
| carvedilol | CV 5 | hypotensives; sympatholytics-beta; vasodilators |
| clonidine | CV 5 | hypotensives; sympathomimetics-alpha |
| atenolol | CV 6 | sympatholytics-beta |
| metoprolol | CV 6 | sympatholytics-beta |
| timolol | CV 6 | sympatholytics-beta |
| famotidine | GI 1 | gastric-secretion-inhibitors; antihistamines-H2 |
| ranitidine | GI 1 | gastric-secretion-inhibitors; antihistamines-H2; antiulcers |
| lansoprazole | GI 2 | gastric-secretion-inhibitors; H−K-atpase-inhibitors |

[a] The two-letter classification codes (AB, AI, CN, CV, GI, and OT) are defined in Figure 7. The numeric identifiers define subclasses within each main class.

are greater than the rotation-dependent scores indicating that these are less sensitive scoring functions. However, in both cases the distance functions flatten after $L = 3$, and it can be seen that RI comparisons with $L \geq 3$ are sufficient to provide a reliable ordering of the database molecules with respect to the query molecule. This suggests that RI comparisons could provide a fast filter to distinguish different molecular shapes rapidly.

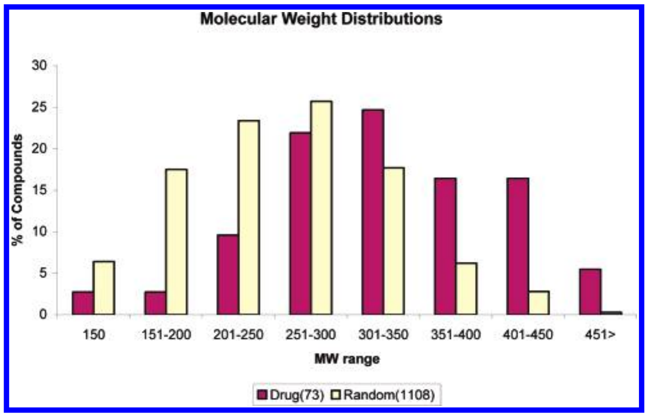

**Figure 1.** Molecular weight distributions of the Drug and Random data sets.

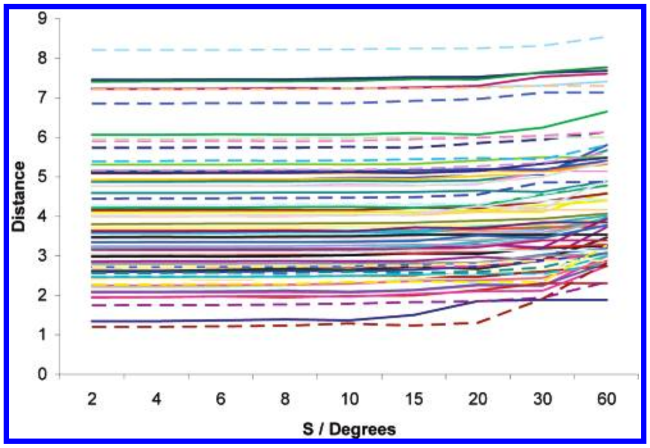

**Figure 2.** Plots of rotation-dependent distance scores (eq 4) as a function of the rotational search angular step size. Each curve represents the calculated minimum distance between lorazepam and one of the remaining 72 molecules in the Drug data set using $L = 6$.

**Database Search**. Because it is relatively time-consuming to calculate high order comparisons, different rotational and RI parameters were tested in order to explore the extent to which the speed of queries on large databases can be improved while still performing accurate searches. Figure 4 shows the resulting ROC curves obtained for a range of expansion orders and search angles when searching the Drug plus Random database using lorazepam as the query molecule. These results show that even low-resolution rotational searches using $L = 2$ and a 15° step size give high TPRs of 80−100% with correspondingly low FPRs of 7−13%, respectively. With expansions between $L = 3$ and $L = 6$, the results improve to 100% TPR with FPRs between 1% and 3%. Retrieval times per match range from 34 ms (ms) for $L = 3$ to 1 s for $L = 6$. Figure 4 also shows the corresponding results for RI expansions from $L = 2$ to $L = 15$. As expected, neither of the RI scoring functions is as accurate as the rotational search. For example, the RIF score recovers around 50% of TPs with a FPR of 9%, although for a higher TPR of 92%, the corresponding FPR increases to 28%. However, comparison times are considerably faster at ∼10 ms/match. RIM scores are essentially independent of the expansion order, being dominated by the $L = 0$ term, and are thus considerably less sensitive than RIF scores. For the ROT and RIF scoring functions, it can be seen that using expansion orders above $L = 6$ give little or no improvement. It can therefore be concluded that scoring with $L = 6$ gives the best balance between efficiency and accuracy.

Although the RI functions may be calculated rapidly, they are significantly less accurate than the more expensive rotational correlation searches. Therefore, the notion of comparing molecular shapes in standard, or canonicalized, orientations was investigated as a way of retaining much of the precision of the rotational $L = 6$ comparisons while avoiding the computational expense of a full rotational search. Hence, each molecule in the Drug plus Random data set was preoriented by rotating their $L = 6$ expansions such that the largest radial extent was aligned with the global *z*-axis and by then applying a pure *z*-axis rotation to place the maximal equatorial extent on the positive *x*-axis. This procedure is similar to aligning the moments of inertia with the principal axes,[43] but using expansions with $L > 2$ eliminates any ambiguity with respect to 180° axis flips. Figure 5 illustrates the canonical orientations of four benzodiazepines. Figure 6 shows the results obtained when querying a database of canonicalized orientations with a canonicalized query molecule in which all distances are calculated using eq 4. This figure shows that canonical comparisons are significantly more accurate than the RI functions and are almost as accurate as full rotational comparisons, giving TPRs of from 80% to 90% with FPRs of only 10−30%. The calculation times for canonical comparisons are essentially identical to those of the RI functions (i.e., a few ms).

**Clustering the Drug and Odor Data Sets**. In order to compare our SH surface comparison approach with other traditional molecular similarity measures, cluster analysis of both the Drug and Odor data sets was performed, and results were compared with PC based clustering of the Drug data set and with the vibrational frequency based clustering of the Odor data set. The Drug data set was initially classified into six broad pharmacological categories and up to seven subgroups based on pharmacological mechanism of action (Table 1) to give a total of 22 drug classes. However, it should be noted that this classification is not unique because many of these compounds have multiple modes of action and are used for a variety of therapeutic purposes. Nonetheless, the classification in Table 1 does provide an indication of pharmacological similarity against which the calculated clusters may be compared. Takane and Mitchell originally clustered the Odor data set into ten distinct groups using eigenvalue (EVA) descriptors derived from vibrational frequency calculations, and the same number of clusters was used in the present study to facilitate comparison with the SH results.

For the PC clustering, 11 molecular descriptors (including polarizibility, radius of gyration, molecular weight, logP, etc.) were calculated for the Drug data set using Cerius-2,[44] and these were autoscaled and clustered using Ward's agglomerative clustering algorithm[45] to produce a total of 22 clusters, as shown in Figure 7. For the SH clustering, a shape-only distance matrix for the same group of molecules using $L = 6$ expansions was calculated, and Ward's algorithm was applied directly to produce 22 clusters, which are also shown in Figure 7. This figure shows that both clustering methods often group similar classes of drugs into the same or similar clusters. For example, both the PC and SH clustering approaches group the antibiotics (AB) together, and both

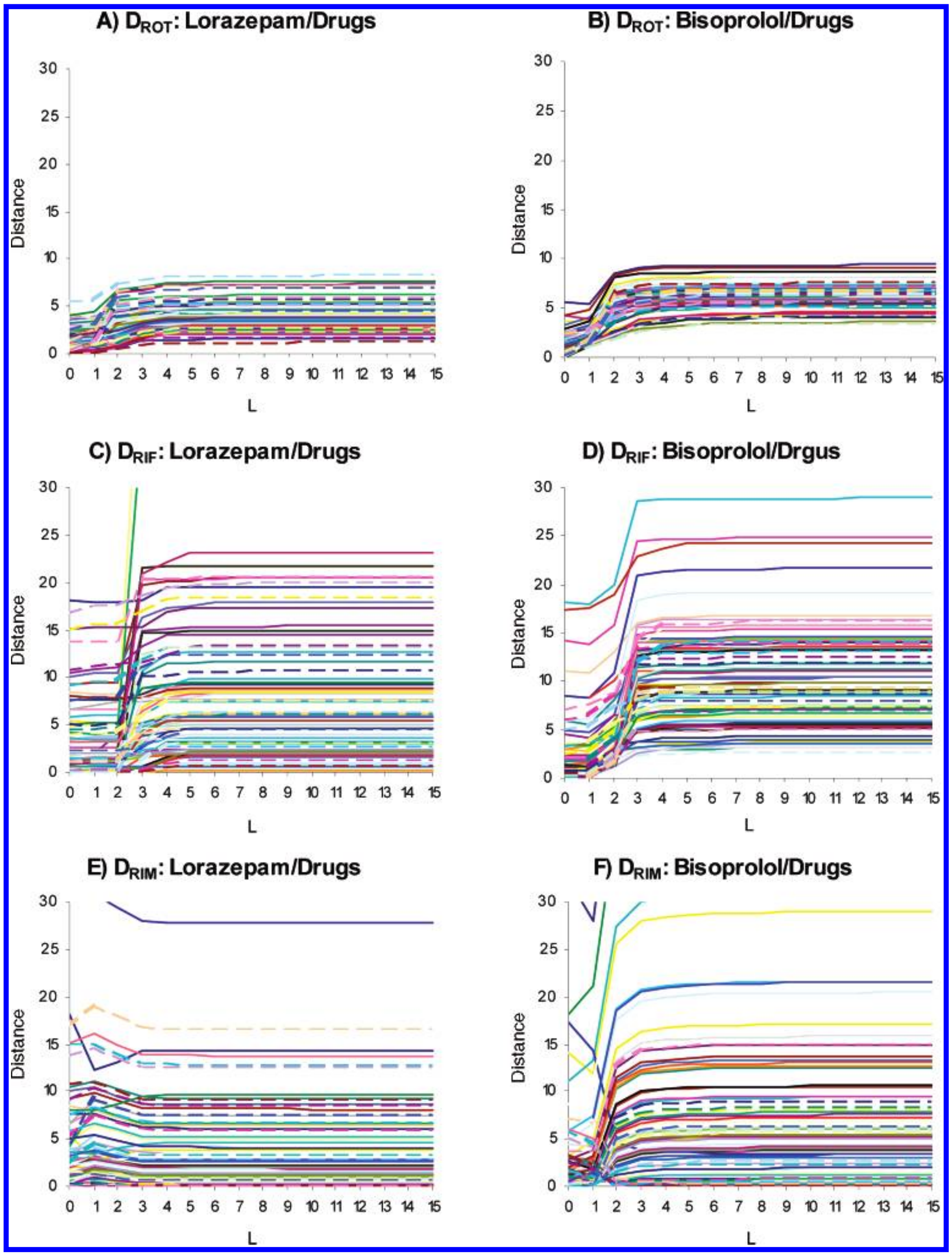

**Figure 3.** Plots of rotation-dependent distances (eq 4) between (A) lorazepam (left), and (B) bisoprolol (right), calculated for each of the remaining 72 drug molecules as a function of the expansion order. C, D: the corresponding plots for the RIF scoring function (eq 7); E, F: the corresponding plots for the RIM scoring function (eq 8).

approaches group many of the tranquilizer and antidepressant drugs (central nervous system; CN) closely together. This would suggest that classifying molecules using SH surfaces is at least as good as traditional methods based on overall molecular properties. Indeed, close comparison of the two dendrograms suggests that the SH clustering tends to place more of the pharmacologically related molecules into more closely related groups than the PC clustering. For example, the SH clustering places the gastrointestinal (GI) drugs in the same group, whereas these drugs are distributed over four distinct groups in the PC clustering. For the CN compounds, one group contains the benzodiazepines clonazepam, lorazepam, and diazepam. The second main CN group primarily contains compounds related to GPCR activity such as the serotonin reuptake inhibitors amitryptiline, nortryptiline, citalopram, fluoxetine, and paroxetine as well as the serotonin receptor antagonist olanzapine. These features of the cluster analysis are not conclusive but are nevertheless encouraging.

As a final test of the SH approach, the Odor data set was clustered into ten groups using SH shape descriptors to $L = 6$. Figure 8 shows the resulting SH clusters along with the corresponding 3D superpositions. Both methods give broadly similar groupings. However, the SH clustering nicely distinguishes the camphoraceous and bitter almond molecules as two separate groups, whereas the earlier EVA clustering study splits the camphors into two subgroups, one of which includes one jasmine and two rose odors (see Table 3 of ref 41). The EVA clustering also splits the bitter almond odors into three distinct groups, whereas the SH clustering correctly assigns these molecules to two neighboring subgroups. The SH clustering also locates all but one of the rose and jasmine

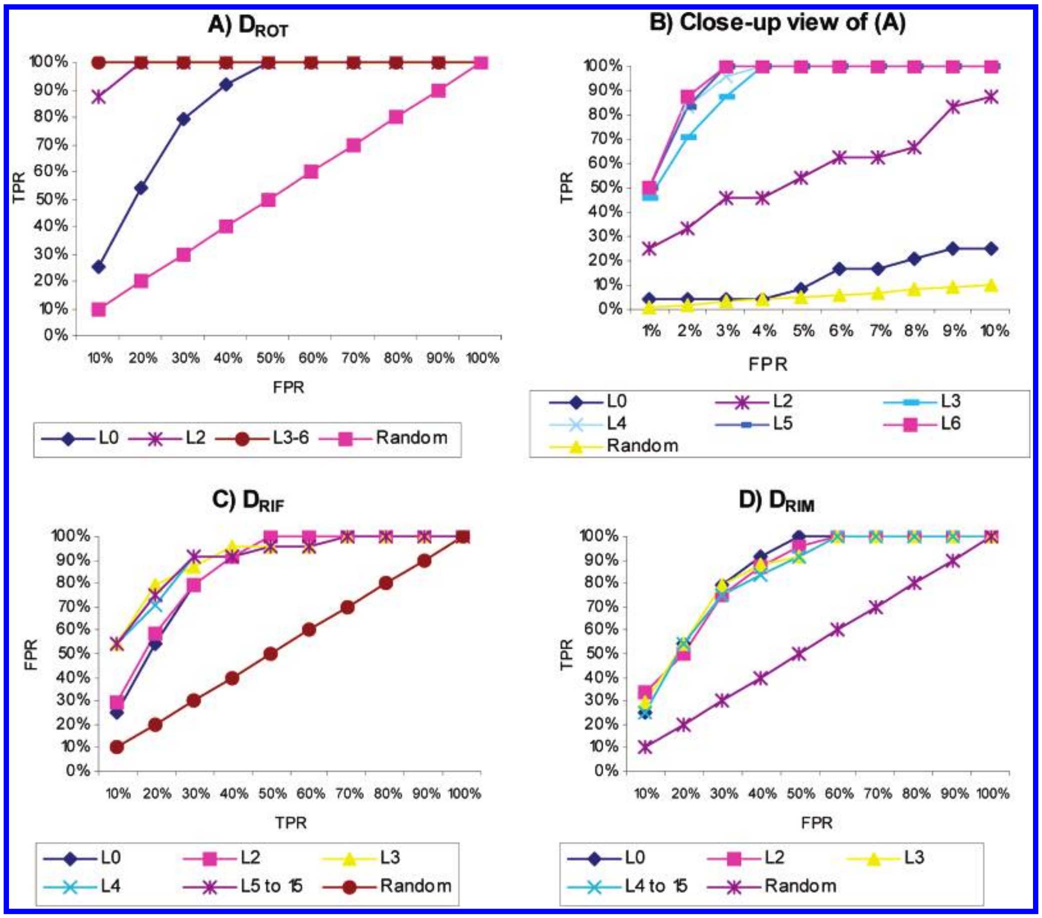

**Figure 4.** ROC plots of database query results for lorazepam and the Drug plus Random data set. A: rotation-dependent scoring function (eq 4); B: close-up view of (A); C: RIF scoring function (eq 7); D: RIM scoring function (eq 8).

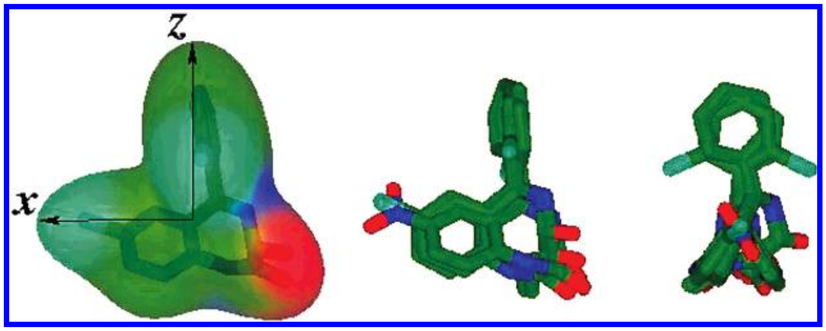

**Figure 5.** Illustration of the canonical alignment of four benzodiazepine GABA receptor agonists: lorazepam, diazepam, temazepam, and clonazepam. Left: the $L = 6$ SH molecular surface and canonicalized orientation of lorazepam. Middle: the four canonicalized benzodiazepines together. Right: the same orientations rotated by 90° about the $z$-axis.

odors in two closely related subgroups. Overall, Figure 8 shows a striking correspondence between the SH shape-based classification and the corresponding molecular shape superpositions.

## DISCUSSION

It has been shown that low-resolution SH expansions provide a reliable and fast way to superpose and compare molecular shapes. In order to accelerate further the scoring calculation, the use of two simple RI scoring functions was investigated. Our database search results show that although the RIF and RIM scores are significantly faster to calculate, they are considerably less precise than rotational searches. Nonetheless, these scoring functions could still provide useful initial filters when searching very large molecular databases.

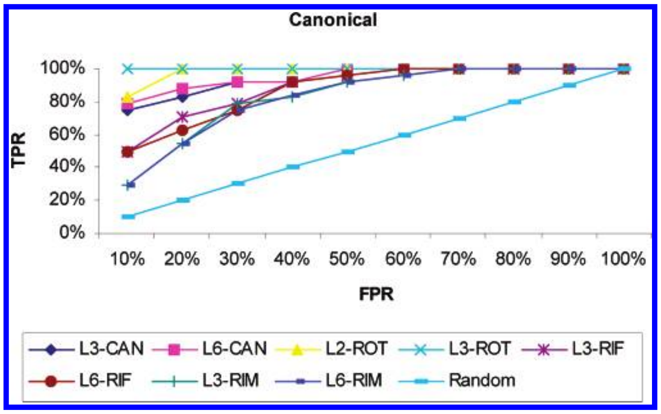

**Figure 6.** ROC plots of rotation-dependent (ROT), rotation-invariant (RIF and RIM), and canonical (CAN) comparison results using lorazepam to query the Drug plus Random data set.

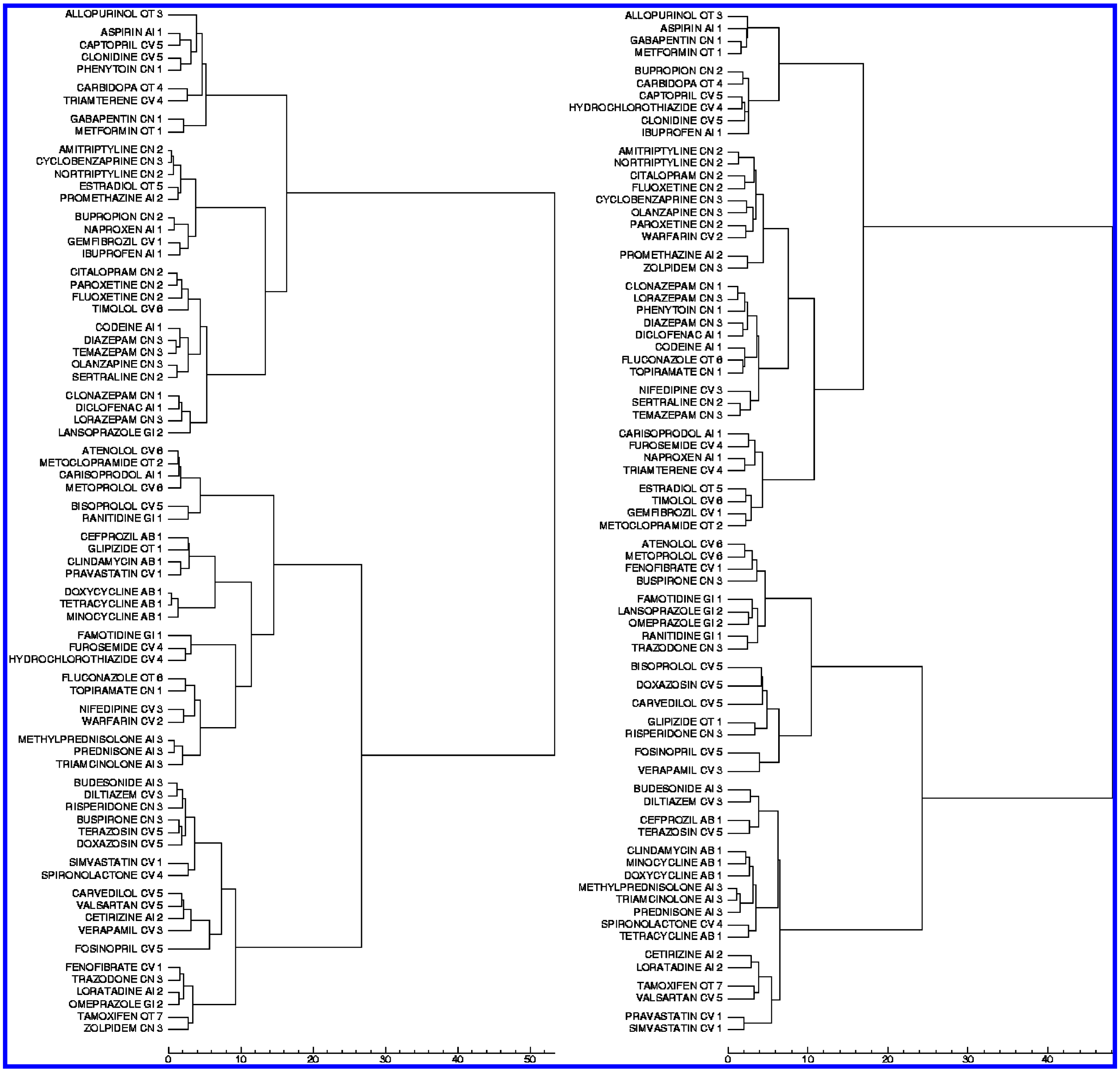

**Figure 7.** Dendrogram of the Drug data set using Ward's agglomerative clustering algorithm to give 22 clusters. Left: conventional chemical clustering using 11 autoscaled macroscopic PC descriptors. Right: $L = 6$ SH surface shape clustering. Each drug is assigned a two-letter code according to its main pharmacological type as follows: AB, antibiotic; AI, anti-inflammatory; CN, central nervous system; CV, cardiovascular; GI, gastrointestinal; OT, other. Numeric identifiers define subclasses within each of the six main pharmacological classes.

Funkhauser et al.[32] and Kazhdan et al.[46] successfully used similar SH-based RI comparisons for the recognition and classification of diverse computer graphics objects. However, our results show that distinguishing and classifying broadly similar shapes such as small ligand molecules require the use of more sensitive orientation-dependent scoring functions. Nonetheless, comparing molecules in precalculated canonical orientations using the full rotation-dependent scoring function was found to give very good sensitivity without sacrificing computational efficiency. Indeed, the computational cost of calculating canonical distance scores is essentially identical to the cost of the RIF scoring function. One minor drawback of using canonical comparisons is that each molecule's SH expansion must be preoriented with respect to the coordinate axes, but this can be done just once when populating the database.

The clustering results of both the Drug and Odor data sets show that SH shape comparisons often give chemically meaningful groupings. Our results for the Odor data set show a striking correspondence between the SH-based classification and the corresponding molecular shape superpositions. Indeed, SH shape clustering achieves comparable or better clusters than previous work based on more computationally expensive quantum mechanics-based vibrational frequency analysis. The analysis of the Drug data set is also encouraging in that, despite using only shape expansions, it is possible to identify similar pharmacological groupings which appear to be at least as good as those produced using traditional

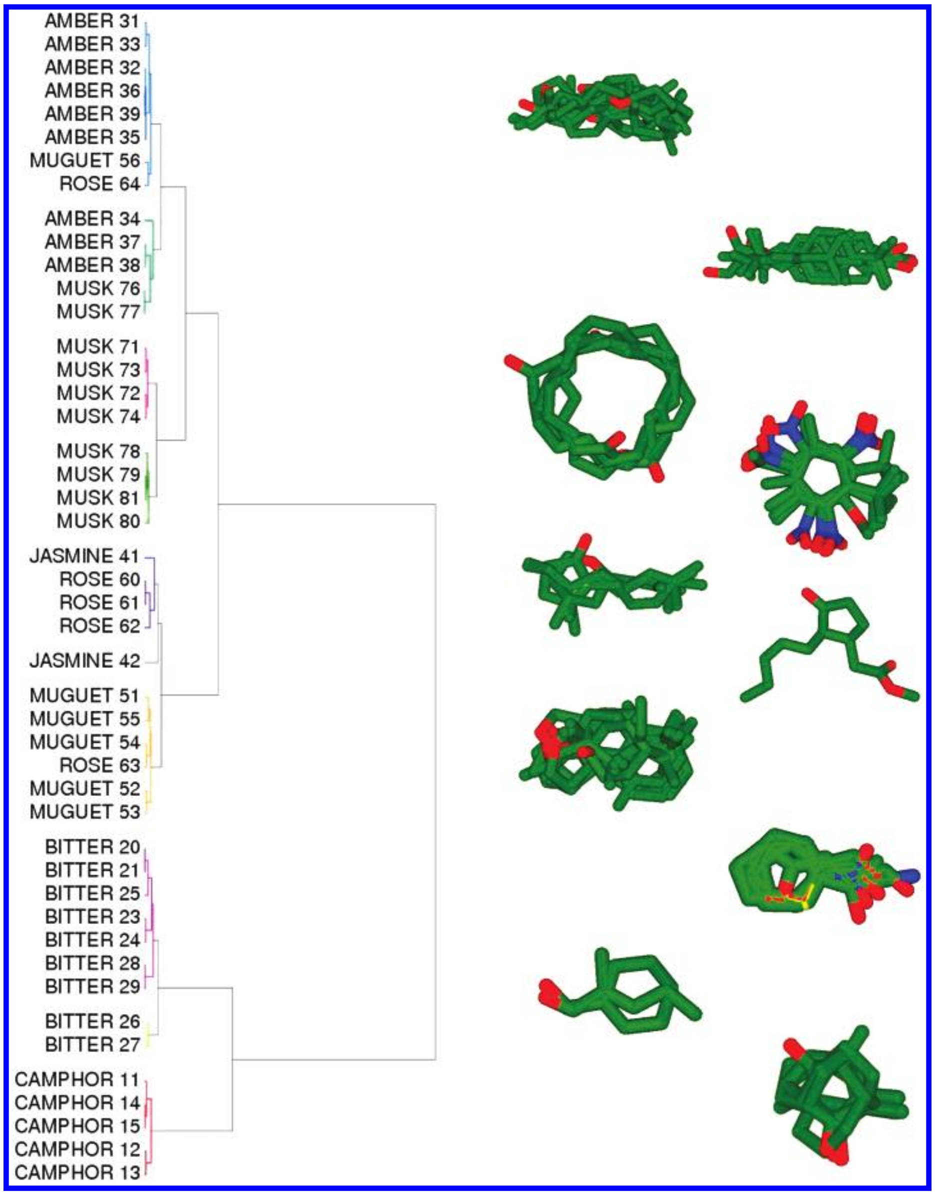

**Figure 8.** SH canonical surface shape clustering of the Odor data set. Left: the dendrogram obtained using $L = 6$ with ten clusters. Right: the corresponding 3D molecular superpositions.

physicochemical based analysis. Because SH shape comparisons are independent of the underlying covalent structures, this suggests that SH shape matching may provide a useful way to formulate scaffold-hopping queries to find novel drug leads.

Finally, it is worth noting that in addition to representing molecular shape, the ParaSurf program of Lin and Clark[47] uses SH expansions to represent other molecular surface properties, including the molecular electrostatic potential, local ionization energy and electron affinity (which are related to local chemical reactivity), and the local polarizability. Indeed, any property that can be described by a surface integral model,[48] e.g. the local contribution to logP, can in principle be represented using SH expansions. Comparing SH property expansions as well as SH surfaces might therefore provide a more specific and sensitive way to formulate 3D scaffold-hopping database queries. We are collaborating to develop ParaFit,[49] a highly optimized C program which implements very rapid SH search, superposition, and clustering of arbitrary combinations of the above molecular surface properties, thus allowing for the first time complex medicinal chemistry based searches to be applied to large 3D data sets. As a simple example, ParaFit permits searching for molecules with both a similar shape *and* similar distributions of hydrophobic/hydrophilic regions: such searches have until now been impossible using existing HTVS techniques. Details and results of ParaSurf and ParaFit will be presented in a subsequent publication.

## CONCLUSIONS

It has been shown that the SH representation provides a powerful way to represent and compare small molecules rapidly and accurately. Our clustering results show that chemically meaningful clusters may be obtained using low order SH expansions. Our database search results show that RI comparisons can provide a fairly crude but very fast initial filter. High accuracy superpositions may be achieved using relatively low order SH rotational correlations. Canonical comparisons are nearly as accurate as full rotational searches but have almost negligible computational cost. These results indicate that SH shape-based correlation techniques could provide a practical and efficient way to search rapidly very large 3D molecular databases. The approach described is currently being extended to allow the rapid calculation and comparison of arbitrary combinations of SH molecular surface properties.

## ACKNOWLEDGMENT

L.M. is funded in part by EPSRC grant no. GR/R72723/1. B.D.H. is supported by BBSRC grant no. BB/E525985/1. We thank Vishwesh Venkatraman for assistance with the preparation of the Odor data set.

## REFERENCES AND NOTES

(1) Lemmen, C.; Lengauer, T. Computational methods for the structural alignment of molecules. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 215−232.

(2) Daylight Chemical Information Systems, Inc. http://www.daylight.com (accessed July 1, 2007).

(3) Tripos Inc. http://www.tripos.com (accessed July 1, 2007).

(4) Elsevier MDL. http://www.mdli.com (accessed July 1, 2007).

(5) Voigt, J. H.; Bienfait, B.; Wang, S.; Nicklaus, M. C. Comparison of the NCI open database with seven large chemical structure databases. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 702−712.

(6) Clark, T. QSAR and QSPR based solely on surface properties? *J. Mol. Graphics Modell.* **2004**, *22*, 519−525.

(7) Brown, N.; Jacoby, E. On Scaffolds and Hopping in Medicinal Chemistry. *Mini-Rev. Med. Chem.* **2006**, *6*, 1217−1229.

(8) Wang, N.; DeLisle, R. K. Fast Small Molecule Similarity Searching with Multiple Alignment Profiles of Molecules Represented in One Dimension. *J. Med. Chem.* **2005**, *48*, 6980−6990.

(9) Haigh, J. A.; Pickup, B. T. Small Molecule Shape-Fingerprints. *J. Chem. Inf. Model.* **2005**, *45*, 673−684.

(10) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A Fast Method of Molecular Shape Comparison: A Simple Application of Gaussian Descriptor of Molecular Shape. *J. Comput. Chem.* **1996**, *17*, 1653−1666.

(11) Cai, W.; Shao, X.; Maigret, B. Protein-Ligand recognition using spherical harmonic molecular surfaces: Towards a fast and efficient filter for large virtual throughput screening. *J. Mol. Graphics Modell.* **2002**, *20*, 313−328.

(12) Morris, R. J.; Najmanovich, R. J.; Kahraman, A.; Thornton, J. M. Real spherical harmonics expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics* **2005**, *21*, 2347−2355.

(13) Tervo, A. J.; Ronkko, T. BRUTUS: Optimization of a Grid-Based Similarity Function for Rigid-Body Molecular Superpositions. 1. Alignment and Virtual Screening Applications. *J. Med. Chem.* **2005**, *48*, 4076−4086.

(14) Hessler, G.; Zimmermann, M. Multiple Ligand-Based Virtual Screening: Methods and Applications of the Mtree Approach. *J. Med. Chem.* **2005**, *48*, 6575−6584.

(15) Yeh, J. S.; Chen, D. Y. A web-based three-dimensional protein retrieval system by matching visual similarity. *Bioinformatics* **2005**, *21*, 3056−3057.

(16) Meurice, N.; Maggiora, G. M. Evaluating molecular similarity using reduced representations of the electron density. *J. Mol. Model.* **2005**, *11*, 237−247.

(17) Rush, T. S., III; Grant, J. A.; Mosyak, L.; Nicholls, A. A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein-Protein Interaction. *J. Med. Chem.* **2005**, *48*, 1489−1495.

(18) Yamagishi, M. E. B.; Martins, N. F.; Neshich, G.; Cai, W.; Shao, X.; Beautrait, A.; Maigret, B. A fast surface-matching procedure for protein-ligand docking. *J. Mol. Model.* **2006**, *12*, 965−972.

(19) *CORINA, version 3.4*; Molecular Networks: Erlangen, Germany, 2006.

(20) Petitjean, M. Geometric Molecular Similarity from Volume-Based Distance Minimization - Application to Saxitoxin and Tetrodotoxin. *J. Comput. Chem.* **1995**, *16*, 80−90.

(21) Hahn, M. Three-Dimensional shape-based searching of conformationally flexible compounds. *J. Comput. Inf. Comput. Sci.* **1997**, *37*, 80−86.

(22) Nissink, J. W.; Verdonk, M. L.; Kroon, J.; Mietzner, T.; Klebe, G. Superposition of molecules: Electron density fitting by application of Fourier transforms. *J. Comput. Chem.* **1997**, *18*, 638−645.

(23) Kearsley, S. K.; Smith, G. M. An Alternative Method for the Alignment of molecular Structures: Maximizing Electrostatic and Steric Overlap. *Tetrahedron Comput. Methodologies* **1990**, *3*, 615−635.

(24) Cruciani, G.; Crivori, P.; Carrupt, P. A.; Testa, B. Molecular fields in quantitative structure-permeation relationships: the VolSurf approach. *J. Mol. Struct.* **2000**, *503*, 17−30.

(25) Goldman, B. B.; Wipke, W. T. QSD quadratic shape descriptors. 2. Molecular docking using quadratic shape descriptors (QSDock). *Proteins: Struct., Funct., Genet.* **2000**, *38*, 79−94.

(26) Platt, D. E.; Silverman, B. D. Registration, Orientation, and Similarity of Molecular Electrostatic Potentials Through Multipole Matching. *J. Comput. Chem.* **1996**, *17*, 358−366.

(27) Good, A. C.; Hodgkin, E. E.; Richards, W. G. The utilisation of Gaussian functions for the rapid evaluation of molecular similarity. *J. Comput. Inf. Comput. Sci.* **1992**, *32*, 188−191.

(28) Boys, S. F. Electronic wave functions I. A general method of calculation for the stationary states of any molecular system. *Proc. R. Soc. London, Ser. A* **1950**, *A200*, 542−554.

(29) Silverman, B. D. Three-dimensional moments of molecular property fields. *J. Comput. Inf. Comput. Sci.* **2000**, *40*, 1470−1476.

(30) Ritchie, D. W.; Kemp, G. J. L. Protein Docking Using Spherical Polar Fourier Correlations. *Proteins: Struct., Funct., Genet.* **2000**, *39*, 178−194.

(31) Ritchie, D. W.; Kemp, G. J. L. Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces. *J. Comput. Chem.* **1999**, *20*, 383−395.

(32) Funkhouser, T.; Min, P.; Kazhdan, M.; Chen, D. Y.; Halderman, A.; Dobkin, D. A Search Engine for 3D Models. *ACM Trans. Graphics* **2003**, *22*, 83−105.

(33) Huang, H.; Shen, L.; Zhang, R.; Makedon, F.; Hettleman, B.; Pearlman, J. Surface alignment of 3D spherical harmonic models: Application to cardiac MRI analysis. *Lecture Notes in Computer Science 3749 - Medical Image Computing and Computer-Assisted Intervention*; 2005; Vol. 8, pp 67−74.

(34) Edvardson, H.; Smedby, O. Compact and efficient 3D shape description through radial function approximation. *Comput. Methods Programs Biomed.* **2003**, *72*, 89−97.

(35) Kovacs, J. A.; Wriggers, W. Fast rotational matching. *Acta Crystallogr. Sect. D: Biol. Crystallogr.* **2002**, *D58*, 1282−1286.

(36) Kahraman, A.; Morris, R. J.; Laskowski, R.; Thornton, J. M. Shape Variation in Protein Binding Pockets and their Ligands. *J. Mol. Biol.* **2007**, *368*, 283−301.

(37) Wigner, E. P. On unitary representations of the inhomogeneous Lorentz group. *Ann. Math.* **1939**, *40*, 149−204.

(38) RxList. http://www.rxlist.com (accessed July 1, 2007).

(39) Chembank. http://chembank.broad.harvard.edu (accessed July 1, 2007).

(40) National Cancer Institute Database. http://cds.dl.ac.uk/cds/datasets/orgchem/isis/nci.html (accessed July 1, 2007).

(41) Takane, S.; Mitchell, J. B. O. A structure-odour relationship study using EVA descriptors and hierarchical clustering. *Org. Biomol. Chem.* **2004**, *2*, 3250−3255.

(42) Egan, J. P. *Signal detection theory and ROC analysis*; Academic Press: New York, 1975.

(43) Lanzavecchia, S.; Cantele, F.; Bellon, P. L. Alignment of 3D structures of macromolecular assemblies. *Bioinformatics* **2001**, *17*, 58−62.

(44) Accelrys Inc. http://www.accelrys.com (accessed July 1, 2007).

(45) Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **1963**, *58*, 236−244.

(46) Kazhdan, M.; Funkhouser, T.; Rusinkiewicz, S. Rotation Invariant Spherical Harmonic Representation of 3D Shape Descriptors. In *Eurographics Symposium on Geometry Processing;* Kobbelt, L., Schroder, P., Hoppe, H., Eds.; Eurographics Association: 2003; pp 156−164.

(47) Lin, J.; Clark, T. An analytical, variable resolution, complete description of static molecules and their intermolecular binding properties. *J. Comput. Inf. Model.* **2005**, *45*, 1010−1016.

(48) Ehresmann, B.; de Groot, M. J.; Alex, A.; Clark, T. New molecular descriptors based on local properties at the molecular surface and a boiling-point model derived from them. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 658−668.

(49) Cepos Insilico Ltd. http://www.ceposinsilico.com (accessed July 1, 2007).

CI7001507