JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

Rapid Shape-Based Ligand Alignment and Virtual Screening Method Based on Atom/Feature-Pair Similarities and Volume Overlap Scoring

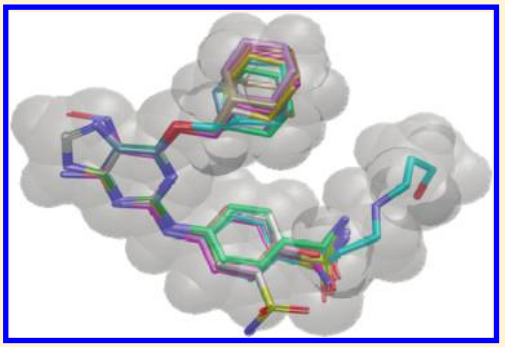
G. Madhavi Sastry,[†,§] Steven L. Dixon,[‡,§] and Woody Sherman*[,‡]

[†]Schrödinger, Sanali Infopark, 8-2-120/113, Banjara Hills, Hyderabad 500034, Andhra Pradesh, India

[‡]Schrödinger, 120 West 45th Street, New York, New York 10036, United States

Supporting Information

ABSTRACT: Shape-based methods for aligning and scoring ligands have proven to be valuable in the field of computer-aided drug design. Here, we describe a new shape-based flexible ligand superposition and virtual screening method, Phase Shape, which is shown to rapidly produce accurate 3D ligand alignments and efficiently enrich actives in virtual screening. We describe the methodology, which is based on the principle of atom distribution triplets to rapidly define trial alignments, followed by refinement of top alignments to maximize the volume overlap. The method can be run in a shape-only mode or it can include atom types or pharmacophore feature encoding, the latter consistently producing the best results for database screening. We apply Phase Shape to flexibly align molecules that bind to the same target and show that the method consistently produces correct alignments when compared with crystal structures. We then illustrate the effectiveness of the method for identifying active compounds in virtual screening of eleven diverse targets. Multiple parameters are explored, including atom typing, query structure conformation, and the database conformer generation protocol. We show that Phase Shape performs well in database screening calculations when compared with other shape-based methods using a common set of actives and decoys from the literature.



INTRODUCTION

Shape-based screening has proven to be a valuable tool in computer-aided drug design, especially in the context of virtual screening.[1−4] Additionally, there have been a number of other successful applications shown in the literature, such as scaffold hopping,[5−7] bioisostere replacement,[8,9] virtual library design,[10] and flexible ligand superposition.[11,12] Shape-based screening, especially when combined with an electrostatic treatment of the atoms, is an attractive method because highly similar molecules in this context should have highly similar binding characteristics. Indeed, shape-based screening methods have proven to perform well in virtual screening calculations when compared with other methods.[13,14] Furthermore, the results have been shown to be complementary with other methods such as docking, and therefore shape-based screening is a valuable tool even when more computationally demanding methods are available.[15] A recent perspective article highlights the role of molecular shape in medicinal chemistry.[16]

In general terms, shape-based virtual screening methods use the shape and other properties of a query molecule to find molecules in a database that resemble the query. Most methods perform an initial 3-dimensional alignment of each database compound to the query molecule, which is followed by a scoring stage where properties in addition to the shape can be considered. A number of methods exist for shape-based screening, all of which have differences in the approach but rely on the same underlying assumption that the shape of a query molecule that is known to be active against a target of interest contains useful information that can help retrieve other active molecules. Among the most widely used shape-based screening methods is ROCS, which was originally developed as a shape only method[17] and later added an electrostatic complementarity term (called ROCS-*color*).[14] ROCS uses a Gaussian representation for atoms, which affords advantages in terms of speed and simplified mathematical operations. Other methods include atom-based clique-matching (SQ/SQW),[12] pharmacophore-based (MedSuMoLig),[18] property-based (USR),[19] geometry-based,[20] and informatics approaches.[21]

All shape-based methods require some way to generate the 3D conformation of the query and screening molecules. When a protein−ligand cocrystal structure is available with the ligand of interest it is typically directly used as the query conformation. Otherwise, a conformational search can be performed on the query ligand and a conformation (typically the lowest in energy) can be used. An adaptive approach has been successfully applied, wherein multiple query conformations were considered and database enrichments were used to guide the choice of query structure.[22] Treatment of screening molecules almost always involves generation of a conformational ensemble, followed by

Received: June 15, 2011
Published: August 28, 2011

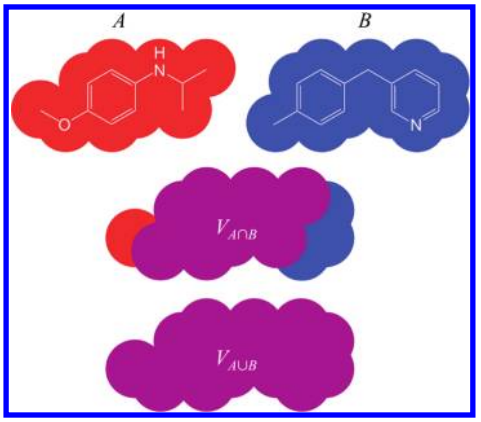


Figure 1. Classical notion of volume based shape similarity between two molecules *A* and *B*, which are represented by sets of atomic spheres. Contributions from heavy atoms only are shown.

selection of the conformation that yields the highest similarity to the query.

While shape-based screening has proven to be valuable, there are a number of drawbacks. First, shape-based screening methods are limited in their ability to find compounds that bind in different modes, different regions of the binding site, or when induced fit effects are important. For example, type II (DFG-out) kinase binders will typically not be found if a type I (DFG-in) ligand is used for the shape query, although this can be accomplished using more sophisticated structure-based approaches[23] that generate induced-fit conformations for protein−ligand complexes.[24] Additionally, as is often the case with ligand-based methods, the results depend heavily on the query molecule. Nonetheless, there is clearly a need for fast and effective methods to find compounds that exhibit similar shape and electrostatic properties to known active compounds.

In this work, we introduce a fast, novel shape-based alignment and screening method, Phase Shape. It relies on atom-pair similarities and rapid overlays of atom triplets. The scoring function can incorporate shape overlap only or a combination of shape and atom/pharmacophore properties. In this paper, we provide details of the methodology, present examples of flexible ligand superposition, apply the method to virtual screening on 11 targets from the MDDR, and compare the results to other methods using the same data.[13]

MATERIALS AND METHODS

Molecular Shape Representation, Alignment, and Similarity. Putta and Beroza[25] outlined the various ways in which molecular shape has been represented historically, with the most widely adopted methods incorporating atomic van der Waals spheres[26] or Gaussian approximations thereof.[27] These volume-based representations are sufficiently rigorous for most purposes and are mathematically convenient for computation of overlaps and similarity. Figure 1 illustrates the classical notion of volume-based shape similarity between two structures *A* and *B*, which are represented by sets of hard atomic spheres. The jointly occupied volume $V_{A\cap B}$ is normalized by the total volume $V_{A\cup B}$ to arrive at a Tanimoto-like shape similarity

$$Sim_{AB} = \frac{V_{A \cap B}}{V_{A \cup B}} \tag{1}$$

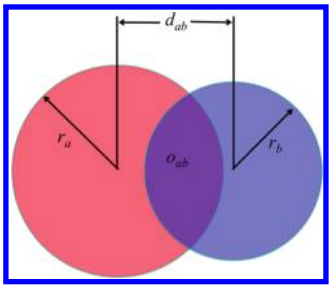


Figure 2. The hard-sphere overlap o_{ab} between atoms *a* and *b*, whose centers are separated by a distance d_{ab}.

This formula is simple and intuitive and is guaranteed to yield a value of one if and only if the shape representations of *A* and *B* are identical, and those shapes have been superimposed correctly.

While eq 1 is simple, it is important to recognize that analytical computation of the hard sphere volumes is not entirely trivial because overlap formulas become increasingly complex as greater numbers of spheres intersect simultaneously.[26] Considerable simplification results when hard spheres are replaced by Gaussian functions,[27] although at the expense of having to compute overlaps between atoms that are separated by far greater than the sum of their van der Waals radii. These issues led us to explore whether a far simpler model of overlap and shape similarity, with a fraction of the computational cost, might lead to a reasonable alternative to eq 1. Figure 2 depicts the elementary case of pairwise overlap between two atomic spheres $a \in A$ and $b \in B$. If the sphere radii are r_a and r_b, and the distance between their centers is d_{ab}, their overlap is given by

$$o_{ab} = \frac{\pi}{12}(r_a + r_b - d_{ab})^2\left[d_{ab} + 2(r_a + r_b) - \frac{3}{d_{ab}}(r_a - r_b)^2\right] \tag{2}$$

Using this model, a total overlap O_{AB} between structures *A* and *B* can be computed by directly summing all pairwise overlaps

$$O_{AB} = \sum_{a \in A}\sum_{b \in B} o_{ab} \tag{3}$$

Because a given element of volume may be counted in more than one pairwise term, O_{AB} is always greater than or equal to $V_{A\cap B}$. Yet O_{AB} still affords a convenient and effective measure of shape consensus,[28] as well as a shape similarity that is restricted to the interval [0, 1] when normalized by the largest self-overlap, where the self-overlap is O_{AA} and O_{BB} for molecules *A* and *B*, respectively

$$Sim_{AB} = \frac{O_{AB}}{\max\{O_{AA}, O_{BB}\}} \tag{4}$$

The denominators in eqs 1 and 4 are somewhat inconsistent, because $V_{A\cup B}$ includes volume from both structures, whereas $\max\{O_{AA}, O_{BB}\}$ includes volume from only one. To address this potential discrepancy, a rigorously computed similarity that is consistent with eq 4 is also introduced

$$Sim_{AB} = \frac{V_{A \cap B}}{\max\{V_A, V_B\}} \tag{5}$$

dx.doi.org/10.1021/ci2002704 |J. Chem. Inf. Model. 2011, 51, 2455–2466

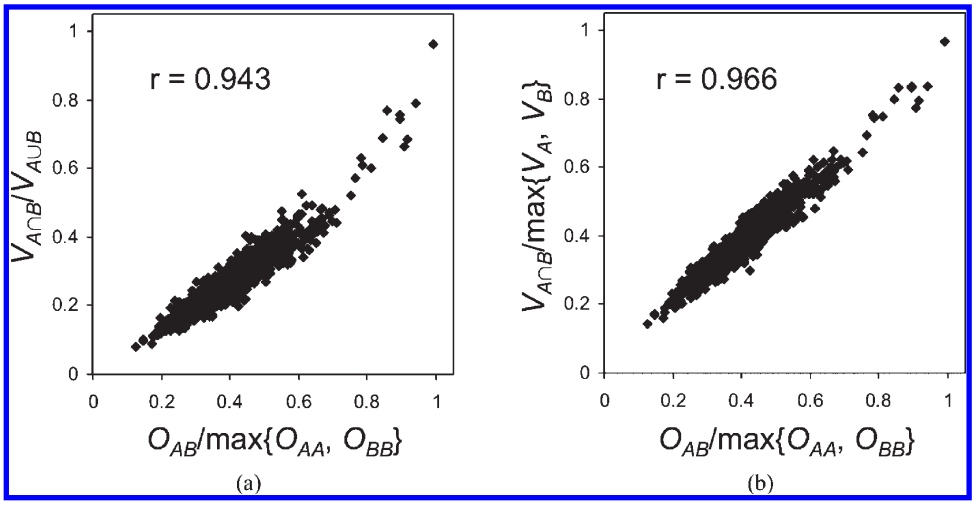


Figure 3. Comparison of shape similarities computed using a rigorous hard sphere volume treatment (*y*-axis) and using sums of pairwise atomic overlaps (*x*-axis). (a) A slightly nonlinear relationship is evident when the rigorous similarity normalization factor is the total volume occupied by both molecules. (b) Linearity is observed when the rigorous and nonrigorous similarities are computed in a consistent fashion, using the maximum individual volume or self-overlap as a normalization factor.

In order to illustrate relationships among the different methods of computing similarity, a set of 50 structures was selected randomly from the Asinex Platinum Collection and aligned to a common query structure using the Phase Shape technology, so that all structures would be in a common frame of reference for subsequent processing. Equations 1, 4, and 5 were then applied to all 1225 unique pairs of structures using a hard sphere model and van der Waals radii. Since each pair was not aligned optimally, the quantities produced were not true shape similarities, but they are still useful for comparing the different approaches.

Figure 3a indicates a strong relationship between the values computed using eqs 1 and 4 (r = 0.943), but a degree of nonlinearity is evident. As shown in Figure 3b, this curvature is eliminated entirely, and the relationship strengthens (r = 0.966) when eq 1 is replaced with eq 5. Thus if one accepts eq 5 as a valid measure of similarity, it is clear that sums of pairwise overlaps provide a satisfactory surrogate for rigorously computed molecular volumes. Furthermore, the computational simplicity of eqs 2 and 3 allows many more overlap calculations per unit time compared to rigorous treatments that employ either hard sphere or Gaussian models.

It is sometimes useful to define a *colored* shape similarity, wherein overlap is counted only between atoms of the same type. Distinguishing by atom type may be critical for eliminating structures that have the desired shape but the wrong electrostatic complementarity. If T comprises the set of distinct atom types in A and B, and A_t and B_t are the atoms of type t in the respective structures, the following generalization of eq 3 affords the appropriate colored overlap

$$O_{AB} = \sum_{t \in T} \left(\sum_{a \in A_t} \sum_{b \in B_t} o_{ab} \right) \quad (6)$$

Whether computing ordinary overlap or colored overlap, obtaining the correct overall shape similarity requires identification of a relative pose of A and B that maximizes O_{AB}. One advantage offered by Gaussian models[27] is that overlaps are smooth, differentiable functions of atomic coordinates, so gradient-based optimization methods can be used to improve a given trial alignment. Hard sphere approaches are not very amenable to such treatments, so in order to find the best alignments, a number of distinct poses of A and B must be investigated. Although a systematic exploration of translations and orientations using a sufficiently fine grid search will yield something very close to the optimal alignment, it is not very practical when rapid shape-based screening is needed. A far more expedient approach is to identify pairs of atoms ($a \in A$, $b \in B$) that have similar local 3D environments and investigate $A \leftrightarrow B$ alignments that are arrived at by superimposing three or more of these pairs. A judicious choice of $a \leftrightarrow b$ atom mappings will invariably lead to alignments that superimpose common structural motifs, and at least one such alignment will be optimal or nearly optimal.

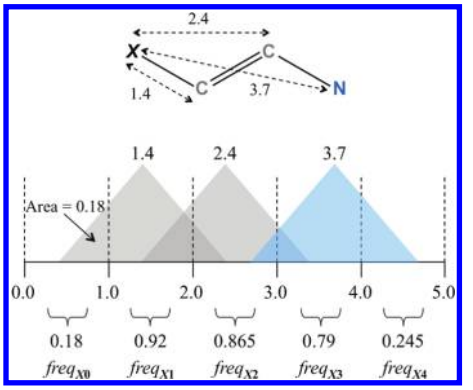


Figure 4. Distribution of atoms according to radial distance from a given atom X. To eliminate edge effects from the binning scheme, each atom is represented by a triangle of unit area, which is divided up among the bins it occupies. Areas from different types of atoms are combined in this example; however, when the similarity calculation distinguishes atom types, each area type is assigned to a distinct set of bins (e.g., bins for X→C, bins for X→N, etc.).

To characterize the local environment of a given atom X, a radial distribution of distances to other atoms is constructed (see Figure 4). Edge effects introduced by the binning scheme are eliminated by representing each atom as a triangle of unit area

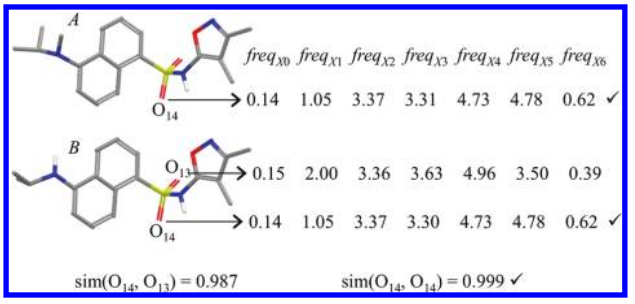


Figure 5. Atom similarities computed as the cosine between distance frequencies over bins 0, 1, ..., 6. In this example, O_{13} and O_{14} in structure *B* are measured to be highly similar to O_{14} in structure *A*, but the higher similarity occurs for the correct pairing, $O_{14}-O_{14}$.

and assigning to a given bin only the fraction of the area that falls into that bin. When colored shape similarity is computed, areas contributed by different atom types are accumulated into distinct sets of bins. So, for example, if the structure in question contains six different types of atoms, the total number of bins is increased by a factor of 6.

If $freq_{ak}$, $k = 0,...,n$ is the distribution of radial distances about an atom $a \in A$, its similarity to an atom $b \in B$ is computed using the following cosine measure

$$\mathrm{sim}(a,b) = \frac{\sum_{k=0}^{n} freq_{ak} \cdot freq_{bk}}{\mathrm{sqrt}\left[\sum_{k=0}^{n} (freq_{ak})^2 \sum_{k=0}^{n} (freq_{bk})^2\right]} \quad (7)$$

By default, $n = 6$, which restricts the similarity information to distances between 0 and 7 Å. This is a good compromise for encoding the unique local environment about a given atom without incurring dominating, highly variable effects from the potentially large numbers of atoms found at longer radial distances. Richmond et al.[29] described a similar procedure for deriving *atom equivalencies*, although they treated atoms as discrete points, and used a total of 20 bins, covering distances up to 20 Å. Furthermore, rather than employing a cosine measure, they defined a cost function c_{ab}, which combined a weighted difference of partial atomic charges q_a and q_b with a sum of scaled, squared histogram differences

$$c_{ab} = w|q_a - q_b| + \sum_{k=1}^{20} \frac{(h_{ak} - h_{bk})^2}{h_{ak} + h_{bk}} \quad (8)$$

Here, h_{ak} is the discrete count of atom centers that lie in the radial distance range $[k-1, k]$ from atom *a*. Figure 5 illustrates application of the cosine similarity defined in eq 7 to the sulfonyl oxygens in a pair of endothelin A antagonists.[30] The local environments of these oxygens are quite similar, and this is borne out by their radial distance distributions. However, the cosine measure assigns a higher similarity to the $O_{14}-O_{14}$ pairing, which is the correct atom mapping between the two structures.

Even when colored similarity is sought, $\mathrm{sim}(a, b)$ does not depend explicitly on the types of *a* and *b*. The best *A*−*B* alignment may very well superimpose atoms that are not of the same type. For example, some of the compounds in the endothelin series from ref 30 contain isoxazole rings in which the oxygen and nitrogen positions are reversed compared to the

b	a_{max}	$\mathrm{sim}(a_{max}, b)$	a_{alt}	$\mathrm{sim}(a_{alt}, b)$
C_{21}	C_{21}	1.000000	C_5	0.966008
O_{14}	O_{14}	0.999999	O_{13}	0.987191
O_{13}	O_{13}	0.999999	O_{14}	0.987260
S_{11}	S_{11}	0.999999	C_6	0.986800
C_{20}	C_{20}	0.999998	C_7	0.991297
C_{19}	C_{19}	0.999997	C_{18}	0.964681
C_{18}	C_{18}	0.999995	C_{19}	0.964678
C_{15}	C_{15}	0.999983	O_{16}	0.986123

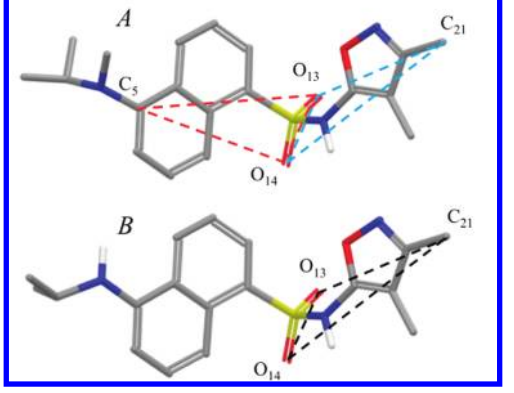


Figure 6. The top eight $B \rightarrow A$ atom mappings and their first alternative mappings. Each set of three mappings yields triads of atoms in the two structures that can be superimposed using a least-squares algorithm to afford a trial alignment of *B* onto *A*.

structures in Figure 5. In such cases, it is important *not* to penalize mappings that involve atom type mismatches because doing so may ultimately lead to a rejection of mappings that would afford the best overall superposition.

Once all pairs of atom similarities have been computed, the following procedure is used to compile a set of $B \rightarrow A$ mappings *SMAP*:

```
for b ∈ B
{
    Find a_max, where sim(a_max, b) ≡ max{sim(a, b): a ∈ A}   (The best mapping for b)
    SMAP{ sim(a_max, b)} → (a_max, b)
}
```

SMAP is a dictionary-style lookup, where the similarity $s(a, b)$ is a key that points to the associated atoms *a* and *b*. Mappings are maintained in order of decreasing similarity, and there is no attempt at this point to ensure that each atom in *B* is mapped to a distinct atom in *A*. Although many of the pairings near the top of *SMAP* will be formally correct, some may not lead to the best overall alignment, so it is important to explore alternative mappings that also exhibit high atom−atom similarities. By default, a single alternative mapping (a_{alt}, b) is identified for a given atom *b*, which comes from the highest value of $\mathrm{sim}(a, b)$, where $a \in A$ and $a \neq a_{max}$. Figure 6 displays the top eight mappings and the corresponding alternatives for the endothelin structures from Figure 5.

Each subset of three mappings yields a pair of triads in the two structures, which can be superimposed using a least-squares procedure to afford a trial alignment of *B* onto *A*. As shown in Figure 6, aligning the black and blue triads from the first three mappings will result in satisfactory overall superposition, whereas

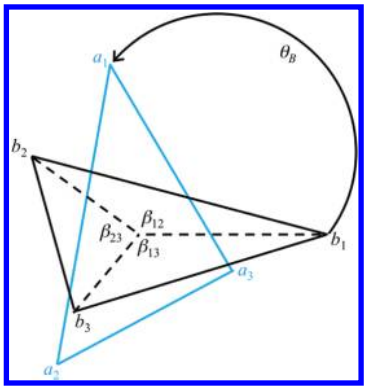


Figure 7. A triad from structure *B* can be rapidly aligned to a triad in structure *A* by shifting both to a common centroid, rotating them to the *xy* plane, and then determining the angle θ_B that minimizes the sum of squared distances between the pairs of atoms (a_1, b_1), (a_2, b_2), and (a_3, b_3).

aligning the black and red triads from the first three alternative mappings will not. The theoretical number of alignments is the product of the number of different triads that can be selected from top n mappings, and the number of ways to construct alternative triads from the m mappings of each atom b

$$\text{Alignments} = \frac{n!}{3!(n-3)!}\,\mathrm{m}^3 \tag{9}$$

With $n = 8$ and $m = 2$, the defaults for Phase Shape, eq 9 yields 448 alignments. This approach can be contrasted with that of Richmond et al.,[29] where a Jonker-Volgenant linear assignment algorithm[31] was used to produce one optimal solution to the atom mapping problem, which may or may not correspond to the best overall superposition of the two structures.

There is no need to perform all 448 triad-based alignments, since many of the triads being compared will be geometrically inconsistent. For example, if a triad in structure *A* is comprised of distances 5, 8, and 7, and it is mapped to a triad in structure *B* with distances 11, 4, and 3, it is clear that the underlying atom mappings cannot possibly be correct, so there is no point in performing the alignment. Phase Shape rejects a triad pairing if any corresponding legs of the triad differ by more than 2 Å, which speeds up the procedure while not rejecting viable candidate triads for the alignment.

For the triads that do have to be aligned, conventional procedures such as reflection-free Procrustes[29] and Ferro-Hermans[32] can be replaced with a significantly faster method that takes advantage of the fact that the least-squares solution will superimpose the triad centroids and place all points in a single plane (i.e., a superposition in 2D), as shown in Figure 7.

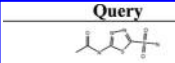
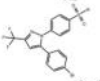
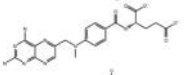
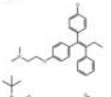
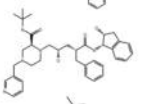
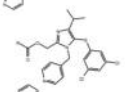
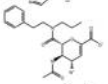
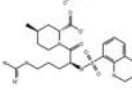
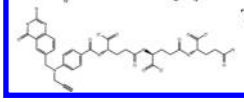
If the centroids are shifted to the origin, and atom b_1 is positioned on the positive *x*-axis, as in Figure 7, the tangent of the optimum angle of rotation θ_B is given by

$$\tan(\theta_B) = \frac{Y}{X} \tag{10}$$

where

$$\begin{aligned} Y \equiv\ & r(b_1)y(a_1) - r(b_2)[x(a_2)\sin(\beta_{12}) - y(a_2)\cos(\beta_{12})] \\ & - r(b_3)[x(a_3)\sin(\beta_{13}) - y(a_3)\cos(\beta_{13})] \end{aligned} \tag{11}$$

Table 1. Systems Studied in This Work, Based on the Original Paper by McGaughey et al.[13]

Query	Target	PDB Code	# Actives
	Carbonic Anhydrase I (CA)	1azm	80
	Cyclin-dependent Kinase 2 (CDK2)	1aq1	77
	Cyclooxygenase 2 (COX2)	1cx2	257
	Dihydrofolate Reductase (DHFR)	3dfr	26
	Estrogen Receptor Alpha (ER)	3ert	74
	HIV Protease (HIVpr)	1hsh	136
	HIV Reverse Transcriptase (HIVrt)	1ep4	149
	Neuraminidase (NA)	1a4q	12
	Protein Tyrosine Phosphatase 1B (PTP1B)	1c87	8
	Thrombin (Throm)	1dwc	200
	Thymidylate Synthase (TS)	2bbq	31

$$\begin{aligned} X \equiv\ & r(b_1)x(a_1) + r(b_2)[x(a_2)\cos(\beta_{12}) + y(a_2)\sin(\beta_{12})] \\ & + r(b_3)[x(a_3)\cos(\beta_{13}) + y(a_3)\sin(\beta_{13})] \end{aligned} \tag{12}$$

Here, $r(b_1)$ is the distance from the origin to atom b_1, $x(a_1)$ is the *x* coordinate of atom a_1, etc. This method requires only about 1/4 the time of a standard least-squares alignment, which leads to a nontrivial reduction in the overall computational time in Phase Shape, since so many alignments are performed.

Structure *B* is subjected to the transformation found using this technique followed by the *inverse* of the transformation that was used to center the structure *A* triad at the origin and rotate it into the *xy* plane. This sequence of transformations results in a superposition of *B* onto *A*, where the two triads are optimally aligned. The total overlap between the two structures is computed using eq 3 or eq 6, and the process is repeated for other triads. After determining the triad-based alignment that yields the highest overlap, a refinement procedure is performed, where additional atoms from *A* and *B* that are within 0.5 Å of being superimposed are identified. A realignment of *B* onto *A* is then performed by applying a conventional least-squares technique to the expanded set of atoms, and this refined alignment is retained if and only if it results in greater total overlap.

Target Ligand Set. The 11 targets from McGaughey et al.[13] were used, and ligand queries were extracted from the appropriate PDB structures (see Table 1). In searching the MDDR for additional actives of each target, the authors of ref 13 encountered

difficulties for CDK2, neuraminidase, and PTP1B. In the case of CDK2, the "Activity Type" field contained nothing specific about CDK2, so the authors used "Protein Kinase C Inhibitor" as a starting point. For neuraminidase and PTP1B, actives could not be found in the MDDR so the authors used a combination of similarity and other keyword searches to find compounds that were considered likely to be actives. This produced only 12 and 8 compounds for neuraminidase and PTP1B, respectively, which were the two smallest actives totals among all 11 targets. While the uncertainties and the small numbers of actives are potentially concerning, we have retained the full set of 11 targets to allow for direct comparison with other methods. Furthermore, average results for the 8-target subset (eliminating CDK2, neuraminidase, and PTP1B) do not change significantly.

The targets were originally chosen by McGaughey et al.[13] based on having at least one high-resolution crystal structure, a large number of structurally diverse active compounds in the MDDR, the inclusion of only a single representative target for a given enzyme family, and spanning a diverse set of active sites (i.e., hydrophobic, hydrophilic, small, large, etc.). The active compounds were taken from the MDDR using queries for the target names. All ligands were prepared using LigPrep.[33]

Database Compounds. The database compounds were taken from the MDDR, as described in McGaughey et al.[13] The initial database of approximately 129,000 compounds was clustered using the Butina algorithm[34] with a similarity cutoff of 0.7 using the Dice similarity metric and atom pair descriptors. The centroid was chosen as the representative structure from each cluster. Molecules with greater than 80 non-hydrogen atoms were removed, as was done by McGaughey et al., resulting in 24,118 compounds for the final decoy set used in this work.

Conformation Generation. Conformational search calculations were performed with the program ConfGen[35] and the utility phasedb_confsites using the 'Rapid' mode. This involves a torsional and ring sampling procedure applied to a core region of the molecule, followed by sampling of peripheral groups, one at a time. Although a 'Thorough' mode exists for sampling peripheral groups simultaneously, we found Rapid sampling to be satisfactory for Phase Shape screens. After generating conformational samples, a soft nonbonded potential is applied to rapidly eliminate unreasonable structures.

A detailed analysis of the effects of conformational search parameters on Phase Shape and other ligand-based methods will be the subject of a separate paper, but as a cursory exploration, we generated four separate Phase conformer databases with varying numbers of conformers per molecule (maximum of 1, 10, 100, and 1000). The average enrichment results improve in going from 1 to 10 to 100, but a plateau is reached in going from 100 and 1000 (see Supporting Information Figure S1). Therefore, the results presented in this paper are all based on a maximum of 100 conformers. Additionally, we considered two different conformations of the query molecule. For most of the paper we show results using the query conformation from the crystal structure noted in Table 1, which is consistent with the work done by McGaughey et al.[13] However, in many cases a crystal structure is not known, so we also wanted to use a query generated with no knowledge of the crystal structure. To do this, we first converted the query to a SMILES string in order to eliminate all conformational biases from the crystal structure. We then converted the SMILES string to 3D using LigPrep and performed a ConfGen conformational search using the Fast mode, retaining the lowest energy structure for the query conformation.

Table 2. Description of the Different Atom/Feature Types Used for the Study

name	description
None	all atoms equivalent (shape-only scoring)
QSAR	Phase QSAR atom types
Element	elemental atom types
MMod	MacroModel atom types
Pharm	Phase pharmacophore feature types

Scoring and Enrichment Calculations. Screens were run for each target with the ligand from Table 1 as the query. Shape similarities were computed between the query and each database compound (actives and decoys). Four atom-typing schemes, shown in Table 2, were explored, ranging from the least to most specific: 1) shape-only (i.e., all atoms treated the same) 2) Phase QSAR[36] (hydrophobic, electron withdrawing, H-bond donor, negative ionic, positive ionic, and other), 3) elemental (separate atom type for each element), and 4) MacroModel (contains over 150 unique atom types). An additional set of screens was performed representing each structure as a set of pharmacophoric features rather than atoms. In this treatment, default Phase feature definitions were applied to identify the locations of all pharmacophore sites of the following type: aromatic, hydrophobic, H-bond acceptor, H-bond donor, negative ionic, and positive ionic. Each site was represented by a hard sphere of radius 2 Å, and as with the atom typing schemes, volume overlap scores were only computed between sites of the same type.

For enrichment calculations we report both the enrichment factor (EF) for the top 1% of the database, the diversity-based enrichment factor (DEF),[37] the Boltzmann-enhanced discrimination of the receiver operating characteristic (BEDROC),[38] and the area under the accumulation curve (AUC). We use $\alpha = 160.9$ for the BEDROC calculations, which corresponds to 80% of the BEDROC score being accounted for in the top 1% of the database screen. While there are differences between all of these enrichment metrics, for much of this work we use the EF(1%) metric because it is more common and can be compared directly to results from other papers, even though DEF and BEDROC have some clear advantages. For example, DEF emphasizes simultaneous retrieval of quantity and diversity of actives, which means retrieving 2 very dissimilar compounds could be better than finding 10 nearly identical compounds. On the other hand, BEDROC is better at differentiating methods in the very early part of a database, since a screen of 5 actives that rank 1, 2, 3, 4, and 5 in will score better than if the same compounds are ranked 6, 7, 8, 9, and 10, even though all actives are within the top 1% of the database screen.

It should be noted that the statistical significance of the enrichment results depends heavily on the number of actives and decoys used in the screens, as described by Truchon and Bayly.[38] The average number of actives across the eleven targets studied here is 95, and the decoys used in all screens is 24,118, which is sufficient for a maximum error in the computed BEDROC value of less than 5% for $\alpha = 20$ (a value of $\alpha = 20$ is what Truchon and Bayly recommend for enrichment studies). However, looking at the very early part of a database, as is often desired, would require a larger database of decoys to get the same computed error. For example, to achieve a maximum error of 5% at the $\alpha = 160.9$ level used in this work would require approximately 160,000 decoy compounds. Since a primary objective of

Table 3. RMSD Values for Alignment of CDK2 Ligands[a]

	1h1q	1h1r	1h1s	1ogu	1oi9	1oiu	1oiy	2c6k	2c6m	2g9x
1h1q	0.4	0.5	0.4	0.5	0.5	0.4	0.3	1.1	0.9	0.7
1h1r	1.5	0.2	1.5	1.6	1.6	1.5	1.5	1.9	2.0	0.7
1h1s	0.7	0.8	0.1	0.3	0.8	0.8	0.4	1.0	1.1	0.3
1ogu	0.4	0.7	0.7	0.2	0.6	0.8	0.7	1.4	0.8	0.7
1oi9	0.5	0.5	0.4	0.4	0.3	0.3	0.5	1.1	1.1	0.7
1oiu	0.9	0.9	2.4	2.6	0.7	0.3	2.5	2.6	3.2	1.0
1oiy	0.4	0.4	0.3	0.5	0.5	0.8	0.2	1.0	1.3	1.0
2c6k	1.6	1.8	1.0	1.2	1.1	1.6	1.1	0.1	0.7	1.6
2c6m	1.2	1.1	0.6	1.0	0.9	1.7	1.2	0.5	0.3	1.1
2g9x	3.4	3.8	3.5	3.5	3.4	3.7	3.7	3.5	2.8	1.2
Average	1.1	1.1	1.1	1.2	1.0	1.2	1.2	1.4	1.4	0.9

[a] The in-place RMSD (Å) values are relative to the respective crystal structure conformation. The high RMSD values for the 2g9x ligand alignments (row 10) are a result of the long flexible tail that does not have a counterpart in any other ligands. Choosing this ligand as the query (column 10) leads to low RMSD values for all ligands, confirming that the conserved parts of the molecules align.

this work was to explore many parameters of the method, it would have been prohibitive to use such a large database. Furthermore, to satisfy the additional objective of comparing with existing methods, we chose the to the MDDR set from McGaughey et al., which had already defined the number of actives and decoys. This data set provides a nice balance between achieving a sufficient level of confidence in the results while still being able to achieve the primary objectives of this work. Finally, it should be emphasized that the number of actives vary substantially in this data set, and cases with very few actives, like neuraminidase and PTP1B, will have a significantly larger standard deviation in the computed enrichment factors.

Computational Times. The search speed for Phase Shape is approximately 500 pregenerated conformers per second. Running Phase Shape on the database with a maximum of 100 conformations per ligand took an average of 0.15 s per ligand (6.8 ligands per second). The fastest calculations were with the neuraminidase query (8.1 ligands per second), and the slowest were with the thymidylate synthase query (5.1 ligands per second). All calculations were run on 2.4 GHz AMD Opteron processors.

RESULTS AND DISCUSSION

Flexible Ligand Superposition. To demonstrate the performance of flexibly superimposing compounds on a query, we begin with a Phase Shape of a 10 × 10 matrix of CDK2 inhibitors from the PDB (see Table 3). Each column represents a different crystal structure ligand being treated as a query molecule. The rows represent each molecule being flexibly aligned to the query. The crystal structure ligands for comparison are brought into the same frame of reference by first running the Align Binding Sites tool in Maestro using Cα atoms within 5.0 Å of the ligand. The diagonal values of the matrix represent the ligands being superimposed on themselves and are generally very low but not exactly zero as a result of the conformational search routine not generating a perfect match to the ligand crystal structure conformation. The 2D structure of each ligand is shown in Figure S2 of the Supporting Information. While superimposing these ten ligands does not represent the most challenging example, as many of the ligands have a common core, it does illustrate a common application of shape-based methods to superimpose a set of related ligands with different substituents that bind to the same target.

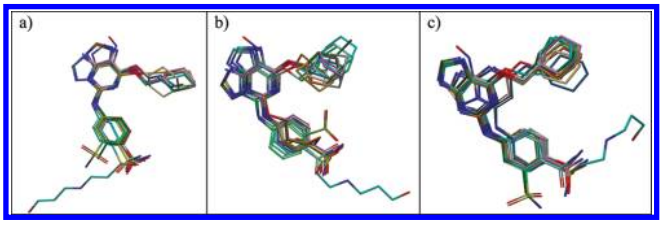


Figure 8. Examples of aligned CDK2 ligands. Heteroatoms of the ligands are colored by element with carbons colored differently for each ligand. The 2g9x ligand is shown in light blue carbons. a) Flexible alignment on the 1h1r ligand; b) Flexible alignment on the 2c6k ligand; c) Crystal ligands aligned by protein binding site, as a reference.

The average RMSD for all 100 flexible superpositions is 1.2 Å and the median is 0.9 Å. The most notable ligands with higher RMSD values are 1oiu and 2g9x. In these two cases, the core of the molecules generally superimpose well with the query (see Figure 8a and b). However, in the case of 2g9x there is a long substituent with six rotatable bonds that extends beyond the shape of any of the other query molecules (see Figure 8c), and the ill-defined placement of this group leads to a high RMSD when aligned to any of the queries except itself. In the case of 1oiu, most of the inhibitors have para-benzyl substitution, whereas the inhibitor in 1oiu has an ortho-sulfonamide substituent, making it difficult to determine where that group should go without knowledge of the receptor. In some cases it goes in the right place (compare Figure 8a and c), whereas in other cases it goes in the wrong place, resulting in a high RMSD (compare Figure 8b and c). The maximum average RMSD for a single query is 1.4 Å (2c6k and 2c6m), and the median for a single query is 1.1 Å. The results presented in Table 3 use MacroModel atom types, which generally yield the best alignments. For comparison, we have included the RMSD matrix for the same set of ligands using shape-only alignments as Table S1 in the Supporting Information. Using shape-only scoring the average RMSD for the full matrix is 1.9 Å, and the maximum average RMSD for a single query is 3.2 Å (2c6k).

To further illustrate the performance of the method for flexible ligand superposition, we provide a few examples from the database screening set. Figure 9a shows the case of a relatively trivial alignment of two COX2 inhibitors where the core of the molecule is superimposed nearly perfectly using the MacroModel atom-typing scheme. This illustrates one of the advantages of Phase Shape, which is that the atom-based alignment algorithm

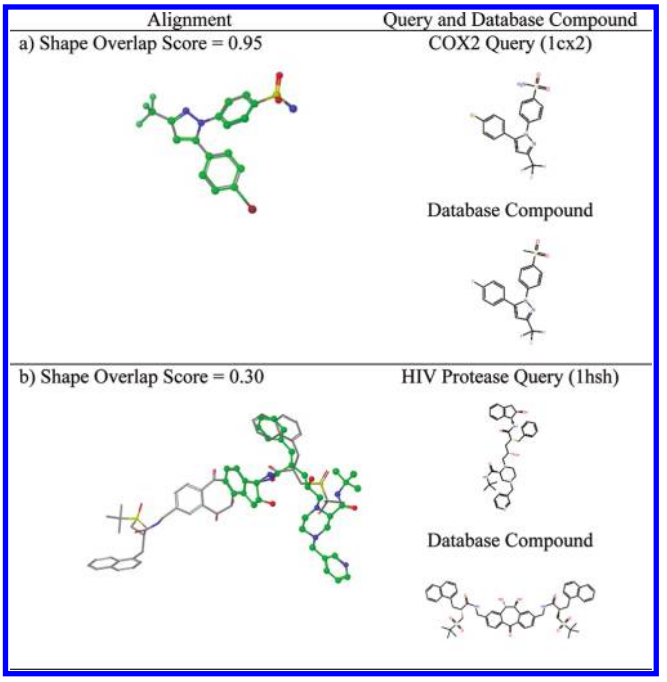


Figure 9. Alignment between the query and a screened compound using MMod atom types. The aligned compounds are shown in tube representation with gray carbons. The query is shown in ball and stick representation with green carbons. a) Active having the best alignment with the query of COX2. The alignment is so close that only one molecule can be seen in most parts. b) An example of a challenging HIV protease query where part of the molecules overlap but large portions of the two molecules do not overlap.

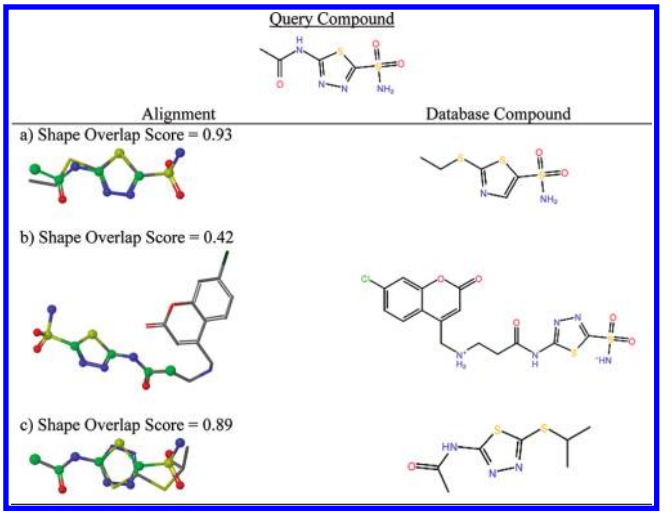


Figure 10. Alignment between the carbonic anhydrase (CA) query and database compounds using shape-only scoring with at most 100 conformers generated for the database compounds. The aligned compounds are shown in tube representation with gray carbons. The query is shown in ball and stick representation with green carbons. Shown are alignment between the query and a) a high ranking active, b) low ranking active, and c) high ranking decoy.

produces excellent alignments of a common core even when R-group substituents vary in shape, size, and chemical functionality. A more difficult example with relatively large and flexible HIV protease inhibitors is shown in Figure 9b. In this case, the crystal structure of only the query is available, so no definite conclusions can be drawn. However, the overall alignment produced here is likely incorrect because the hydroxyl groups on the central cycloheptane ring should be interacting with the catalytic aspartates in the same way as the central hydroxyl in the peptidomimetic query. Despite this discrepancy, it can be seen that some of the rings and peptide bonds in the two structures superimpose very well.

Figure 10 shows a variety of alignments for carbonic anhydrase (CA) inhibitors using shape-only scoring (i.e., no atom types). The query is a small compound with a central thiadiazole and two small substituents (a sulfonamide and an amide). Figure 10a shows a relatively simple alignment of compounds with the same central ring and substituents of similar size. Although this case is relatively easy, it is reassuring to see how precisely the atoms of the core superimpose even though the substituents are different. This demonstrates the advantage of the atom-based Phase Shape alignment methodology. Figure 10b shows an alignment of an active compound with the same central core but a much larger substituent off the amide moiety. While the overall shape overlap score is moderate, the alignment of the common core is excellent, highlighting one of the features of an atom-based approach to shape-based alignment. The alignment of the core is so accurate that it is hard to tell that there are two molecules. Finally, Figure 10c shows an example of a molecule with a high shape overlap score even though the core does not superimpose perfectly. This is likely due to the difference in geometry between the sulfonamide in the query and the isopropyl sulfide in the screening compound. The small degradation in the overlap of the core is compensated by a high overlap of the different substituents.

Finally, Figure 11 shows a series of alignments to a DHFR query molecule. Figure 11a is an example of a similar molecule that receives a relatively high shape overlap score, even though there is a significant difference in the linker chemistry between the two parts of the molecule. Figure 11b shows a more difficult example of a much larger active molecule, which produces a low RMSD alignment for the common part of the molecule and therefore a relatively high overlap score, even though much of the database molecule cannot be matched to any part of the query. Interestingly, Figure 11c shows an example of a high-scoring decoy molecule from the MDDR. While this molecule is not classified as a DHFR inhibitor in the MDDR, it clearly has a high similarity to the query, differing by only a hydroxyl group, and should be retrieved even if the compound is not reported as active or is below the detectable limit of the experimental assay.

Database Virtual Screening. Next, we apply the method to virtual screening and compare the different treatment of atom/feature types (see Table 4). The shape-only method performs the poorest, with an average enrichment across the 11 targets of only 11.9. The relatively poor performance of shape-only scoring is expected because important information about atoms and features is omitted. Our results are consistent with the observations by McGaughey et al., where the shape-only methods (SQW-shape and ROCS) yielded average EF(1%) across the same data set of 8.8 and 9.3, respectively. Application of atom-typing rules to the volume overlap scoring such that volume overlap is only considered between the same atom types improves the results. In fact, as the atom-typing scheme gets progressively more specific, the enrichments consistently increase. The most generic atom typing is Phase QSAR, which produces an average EF(1%) of 15.6. With elemental atom typing, average EF(1%) increases to

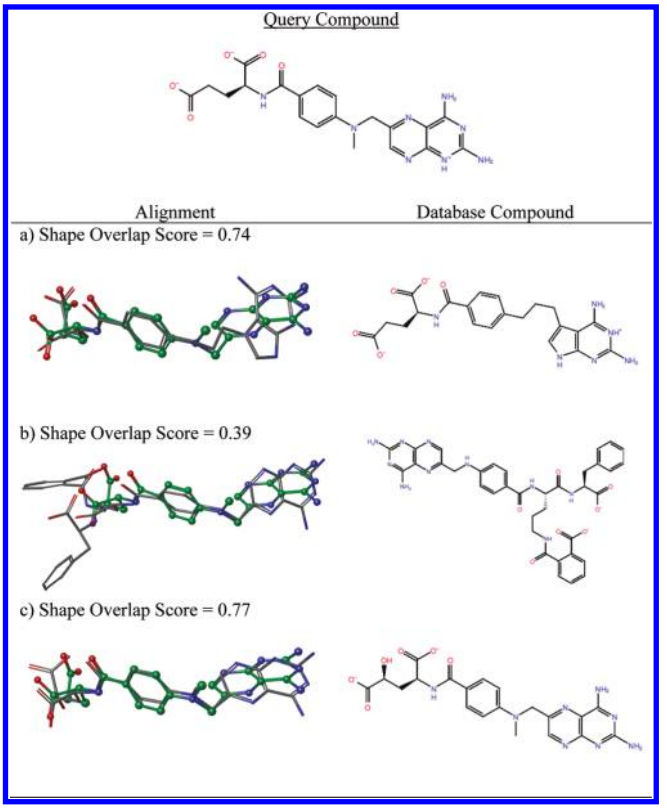


Figure 11. DHFR screen with MMod atom types. The aligned compounds are shown in tube representation with gray carbons. The query is shown in ball and stick representation with green carbons. Alignment between the query and a) high ranking active, b) low ranking active, and c) high ranking decoy.

17.0, and the most specific atom-typing scheme, MacroModel, yields an EF(1%) of 20.0. Interestingly, these incremental improvements are all but overshadowed by the dramatic leap in performance that occurs when pharmacophore feature types are used. In this case, average EF(1%) increases to 33.2, which is the result of five cases getting significantly better (DHFR, ER, NA, PTP1B, and Throm) and only two cases getting slightly worse (CDK2 and HIVrt) than the next best results, which were obtained using MacroModel atom types. Based on the superior performance of the pharmacophore-based approach, we focus on this method while investigating the effect of varying other parameters.

Table 5 shows the comparison of the different enrichment metrics (EF(1%), DEF(1%), BEDROC (α = 160.9), and AUC) results for each of the 11 targets using pharmacophore feature scoring. While there are variations in the details of each metric, the overall trends across the targets are the same. In fact, the differences between the various atom/feature typing for volume scoring also results in qualitatively similar conclusions when looking at the different enrichment metrics (data not shown). Note the comparison of DHFR and TS results in Table 5, where the AUC values are almost identical between these two targets, but EF(1%) shows significantly better performance on DHFR. This effect is magnified when looking at BEDROC (α = 160), with the DHFR number almost 30% higher than the TS value, illustrating the additional emphasis on early enrichment using the BEDROC metric. Indeed, looking at the top 0.1% of the screened compounds we find 10 actives for DHFR (ranked 2, 4, 5, 6, 7, 8, 12, 15, 16, and 21) but only 8 actives for TS (ranked 1, 2, 3, 8, 10, 16, 21, and 22). Regarding the DEF metric, all values except NA and PTP1B are within 10% of the EF values, showing that Phase Shape is retrieving most if not all of the full diversity of the active compounds in the early part of the screen.

Next, we explore the effect of different conformational search methods, both for the database and query molecules using similarities and alignment based on pharmacophore features. When a crystal structure of the query ligand of interest is not known, it is necessary to determine an appropriate 3D conformation for screening. In previous work we showed that ConfGen is able to accurately reproduce the bioactive conformation of drug-like molecules in a large test set.[35] Rather than explore many potential ways of generating query conformations, here we explored the most straightforward approach, which is to use

Table 4. Database Enrichment Performance Using Different Atom/Feature Typing[a]

	EF(1%)				
target	Shape Only	QSAR	Element	MMod	Pharm
CA	10.0	25.0	27.5	32.5	32.5
CDK2	16.9	20.8	20.8	23.4	19.5
COX2	21.4	19.1	16.7	19.5	21.0
DHFR	7.7	3.9	11.5	23.1	80.8
ER	9.5	17.6	17.6	13.5	28.4
HIVpr	13.2	17.7	19.1	14.0	16.9
HIVrt	2.7	2.0	4.7	4.7	2.0
NA	16.7	16.7	16.7	16.7	25.0
PTP1B	12.5	12.5	12.5	12.5	50.0
Throm	1.5	4.0	4.5	8.5	28.0
TS	19.4	32.3	35.5	51.7	61.3
average	**11.9**	**15.6**	**17.0**	**20.0**	**33.2**
median	**12.5**	**17.6**	**16.7**	**16.7**	**28.0**

[a] The search is performed using the crystal conformation for the query and generating maximum of 100 conformers for the database compounds.

Table 5. Database Enrichment Performance Using Pharmacophore Feature Types[a]

target	EF(1%)	DEF(1%)	BEDROC (α = 160.9)	AUC
CA	32.5	32.5	0.29	0.86
CDK2	19.5	19.1	0.19	0.59
COX2	21.0	20.7	0.33	0.70
DHFR	80.8	75.4	0.62	0.96
ER	28.4	27.2	0.25	0.73
HIVpr	16.9	15.3	0.18	0.73
HIVrt	2.0	1.9	0.03	0.55
NA	25.0	19.0	0.30	0.93
PTP1B	50.0	41.0	0.29	0.68
Throm	28.0	27.2	0.31	0.84
TS	61.3	57.7	0.46	0.95
average	**33.2**	**30.6**	**0.30**	**0.78**
median	**28.0**	**27.2**	**0.30**	**0.73**

[a] The search was performed as in Table 4.

Table 6. Comparison of Enrichment Factor (EF(1%)) Using the Crystal Structure or the Lowest Energy ConfGen Conformation for the Query (RMSD Values Å)[a]

target	Crystal	ConfGen	RMSD
CA	32.5	47.5	1.2
CDK2	19.5	20.8	0.3
COX2	21.0	23.0	0.9
DHFR	80.8	53.9	1.7
ER	28.4	23.0	0.9
HIVpr	16.9	9.6	2.1
HIVrt	2.0	2.7	2.4
NA	25.0	33.4	0.7
PTP1B	50.0	50.0	0.3
Throm	28.0	29.5	3.0
TS	61.3	58.1	5.8
average	**33.2**	**31.9**	**1.8**
median	**28.0**	**29.5**	**1.2**

[a] The database compound conformations were generated as in Table 4.

Table 7. Database Enrichment Performance (EF(1%)) Using Pharmacophore Feature Similarity and Two Conformational Search Methods Using the Crystal Structure Query[a]

target	ConfGen Phase	ConfGen Fast CF
CA	32.5	33.8
CDK2	19.5	19.5
COX2	21.0	21.8
DHFR	80.8	80.8
ER	28.4	28.4
HIVpr	16.9	14.7
HIVrt	2.0	2.0
NA	25.0	33.4
PTP1B	50.0	50.0
Thrombin	28.0	27.0
TS	61.3	58.1
average	**33.2**	**33.6**
median	**28.0**	**28.4**

[a] Phase ConfGen is the built-in method in Phase Shape, whereas ConfGen Fast CF uses parameters developed by Chen and Foloppe to optimize ConfGen performance for bioactive conformer retrieval.[39]

the lowest energy conformation from ConfGen using the Fast search mode. As seen in Table 6, the average enrichment degrades only slightly, from 33.2 to 31.9, when using the low-energy ConfGen query, while the median enrichment is actually improved, from 28.0 to 29.5. These results suggest that using the lowest energy conformation from a ConfGen search as a query is a good way to generate high enrichments when a crystal structure query conformation is not available. Table 6 also shows the RMSD value of the low-energy ConfGen conformation compared with the crystal structure. In most cases, the RMSD value between the crystal and ConfGen conformations is below 2.0 Å, which is particularly encouraging given that only a single ConfGen structure was generated. TS yields the highest RMSD, which is not surprising given that the query contains 22 rotatable bonds, the highest among all of the queries. Other highly flexible queries include the ligands for HIVpr and thrombin, which contain 14 and 10 rotatable bonds, respectively. For the cases with RMSD over 2.0 Å, only HIVpr shows a significant decrease in enrichment.

Table 8. Enrichment (EF(1%)) Comparison with Other Shape-Based Screening Methods[a]

target	Phase Shape	ROCS-color	SQW
CA	32.5	31.4	6.3
CDK2	19.5	18.2	9.1
COX2	21.0	25.4	11.3
DHFR	80.8	38.6	46.3
ER	28.4	21.7	23.0
HIVpr	16.9	12.5	5.9
HIVrt	2.0	2.0	5.4
NA	25.0	92.0	25.1
PTP1B	50.0	12.5	50.2
Throm	28.0	21.1	27.1
TS	61.3	6.5	48.5
average	**33.2**	**25.6**	**23.5**
median	**28.0**	**21.1**	**23.0**

[a] Phase Shape uses pharmacophore feature types and the default ConfGen method for generating conformations for database compounds. The ROCS-color and SQW results are obtained from McGaughey et al.[13] and use chemical typing, which produces better results than shape-only screening.

Although the goal of this paper is not an exhaustive exploration of conformational search settings, it is important to determine the sensitivity of the results to the conformational search protocol for the database compounds. As mentioned in the Materials and Methods section and as shown in Figure S1, the enrichments improve with the maximum number of conformations up to 100 and then plateau. However, there are different search protocols that one can use to generate 100 conformations per molecule. Table 7 compares results using the 100-conformer databases generated using the ConfGen Phase protocol and the Fast-CF mode of ConfGen.[39] The Fast-CF method was proposed by Chen and Foloppe to improve the fraction of cases with RMSD below 1.0 Å to the bioactive conformation. As seen in Table 7, the enrichments of the two methods are not significantly different, with averages of 33.2 and 33.6 for ConfGen Phase and Fast-CF, respectively. This result illustrates the robustness of the method to different conformational search methods, although in both cases the methods have been shown to accurately reproduce bioactive conformations. It is possible that a conformational search method that does not accurately and consistently reproduce bioactive conformations would yield degraded Phase Shape enrichments.

Finally, it is useful to compare the results with other methods from the McGaughey et al.[13] paper run on the same data set. The closest methods to Phase Shape, both in terms of performance and methodology, are SQW and ROCS-color. Figure 8 shows the comparison of these three methods using the settings that produce the best results. ROCS-color combines a pure Gaussian-based shape description with a "color" force field that differentiates atoms according to the following types: cations, anions, hydrogen bond donors, hydrogen bond acceptors, and hydrophobes. SQW, a clique-based matching method, incorporates a comparable atom-typing scheme, but similarity between two structures is computed using a Dice measure, and scoring between matching atoms is based not on overlap but on the

similarity of the atom types and the distance between them.[12] We see that ROCS-color produces a mean EF(1%) of 25.6 and a median of 21.1. SQW performs comparably to ROCS-color, with a slightly lower average of 23.5 but a slightly improved median of 23.0. Phase Shape with pharmacophore feature typing performs the best, both in terms of mean and median, with values of 33.2 and 28.0, respectively. Phase Shape performs better than the other two methods on 7 of the 11 targets. In 3 of 4 cases where it does not (COX2, HIVrt, and PTP1B), the differences are less than 5 enrichment units. The largest difference comes from neuraminidase, where ROCS-color significantly outperforms the other methods (See Table 8). Neuraminidase is one of the targets from the original McGaughey et al. paper that did not have known actives in the MDDR, so a combination of similarity and keyword searching was performed to find putative active molecules. With uncertainty in the activity of the compounds and only 12 putative active compounds in total, the statistics are possibly less significant than for the other targets. It should be noted that the comparison presented above uses the most recent version of Phase Shape but older versions of ROCS and SQW, so it is likely that the performance of those programs has improved since the original McGaughey publication.

While the shape-based methods generally performed well in the McGaughey et al.[13] study, the best method was TOPOSIM, which is a 2D similarity method.[40] TOPOSIM had mean and median EF(1%) of 29.0 and 28.6, respectively, which is lower than the Phase Shape mean but slightly higher than the median. Interestingly, in another paper recently published by our group, we obtained average EF(1%) enrichments on this same data set as high as 35.1 using 2D fingerprints,[41] suggesting that fingerprint-based methods also offer a fast way to retrieve a large number of similar compounds. However, a very large number of fingerprint settings were explored in that work to achieve such high enrichments, while Phase Shape with default pharmacophore feature types was able to achieve a high enrichment on par with the best fingerprint results. In general, more sophisticated methods like shape-based screening, pharmacophore modeling, and docking offer complementary approaches to 2D fingerprint methods and often retrieve more diverse compounds,[37] suggesting that using a diversity-based enrichment factor, such as the DEF metric proposed by Salam et al.,[37] could offer a convenient way of augmenting standard enrichment factors to simultaneously account for chemical diversity and enrichment of actives. Unfortunately, it is not possible to perform a direct comparison of the DEF metric for the methods published by McGaughey et al.[13] because we do not have access to the rank of each active compound from that work.

CONCLUSION

We have described a new method called Phase Shape for performing shape-based flexible ligand alignments and database screening. The method is based on an initial alignment of many similar atom triplets followed by a refinement using a more extended set of atoms for the top alignments. Overlap scoring can be based on shape only or a variety of atom/feature typing rules. We find that the atom-based nature of Phase Shape results in accurate ligand superpositions, with the best alignments coming from using the most specific atom types (MacroModel atom types). For database enrichment studies, we find that using atom/feature typing produces a consistent improvement in enrichment relative to shape-only scoring. The pharmacophore-based alignment and scoring scheme produces the best database screening results, with average enrichments significantly higher than the other schemes presented here. We also found that Phase Shape performed favorably when compared to other shape-based methods (SQW and ROCS-color) previously published[13] using this same data set.

Phase Shape was shown to be robust to various parameters explored, including a number of conformations generated for the database compounds, the method for generating database compound conformations, and the query conformation. The latter is particularly important because it represents the scenario where a desired query molecule does not have a crystal structure. In this case, using the lowest energy query conformation from ConfGen yields enrichment results comparable with those obtained using the crystal structure query conformation. In addition, the Phase Shape results were shown to produce high diversity in the enrichment results, as indicated by the diversity-based enrichment factor (DEF) metric.

While the overall results for this method are encouraging, we note that Phase Shape suffers from the same limitations as other shape-based screening methods. For example, a single query molecule in a single conformation is unlikely to contain all of the shape/property information needed to recover all known actives for a given target, especially when substantial conformational rearrangement of the protein (i.e., induced fit) occurs for different ligands. Furthermore, even for a single protein conformation, it is possible for ligands to bind in different modes or interact with different subpockets. In such cases it might be advantageous to use an ensemble query approach, which will be the focus of future studies. Additionally, in future work we will explore in more detail the effects of variations in the conformer generation methods for Phase Shape and other 3D ligand-based screening methods. In this future work we plan to expand the target set, which will improve the statistical significance of the results. The use of the eleven targets presented in this work was necessary to make a direct comparison with other methods, but moving forward it will be important to look at a larger data set with more coverage of pharmaceutically relevant targets. In summary, we find Phase Shape to be an accurate and robust method for shape-based flexible superposition and database screening that can be used to complement other methods, such as 2D fingerprint similarity searching and structure-based docking.

ASSOCIATED CONTENT

Supporting Information. An additional figure showing the database screening enrichment performance for each target using different numbers conformations is provided in Figure S1. Figure S2 shows the 2D representation of the ligands used in the CDK2 flexible superposition study. Table S1 shows the RMSD values for shape-only flexible superposition of the CDK2 ligands. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Phone: (212) 295-5800. Fax: (212) 295-5801. E-mail: Woody.Sherman@schrodinger.com.

Author Contributions

§These authors contributed equally.

REFERENCES

(1) Kirchmair, J.; Distinto, S.; Markt, P.; Schuster, D.; Spitzer, G. M.; Liedl, K. R.; Wolber, G. How To Optimize Shape-Based Virtual Screening: Choosing the Right Query and Including Chemical Information. *J. Chem. Inf. Model.* **2009**, *49*, 678–692.

(2) Singh, J.; Chuaqui, C. E.; Boriack-Sjodin, P.; Lee, W. C.; Pontz, T.; Corbley, M. J.; Cheung, H. K.; Arduini, R. M.; Mead, J. N.; Newman, M. N. Successful shape-based virtual screening: The discovery of a potent inhibitor of the type I TGFβ receptor kinase (TβRI). *Bioorg. Med. Chem. Lett.* **2003**, *13*, 4355–4359.

(3) Noha, S. M.; Atanasov, A. G.; Schuster, D.; Markt, P.; Fakhrudin, N.; Heiss, E. H.; Schrammel, O.; Rollinger, J. M.; Stuppner, H.; Dirsch, V. M.; Wolber, G. Discovery of a novel IKK-β inhibitor by ligand-based virtual screening techniques. *Bioorg. Med. Chem. Lett.* **2011**, *21*, 577–583.

(4) LaLonde, J. M.; Elban, M. A.; Courter, J. R.; Sugawara, A.; Soeta, T.; Madani, N.; Princiotto, A. M.; Kwon, Y. D.; Kwong, P. D.; Schn, A.; Freire, E.; Sodroski, J.; Smith Iii, A. B. Design, synthesis and biological evaluation of small molecule inhibitors of CD4-gp120 binding based on virtual screening. *Bioorg. Med. Chem.* **2011**, *19*, 91–101.

(5) Rush, T. S.; 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.

(6) Oyarzabal, J.; Howe, T.; Alcazar, J.; Andre s, J. I.; Alvarez, R. M.; Dautzenberg, F.; Iturrino, L.; Marti nez, S.; Van der Linden, I. Novel approach for chemotype hopping based on annotated databases of chemically feasible fragments and a prospective case study: new melanin concentrating hormone antagonists. *J. Med. Chem.* **2009**, *52*, 2076–2089.

(7) Boström, J.; Berggren, K.; Elebring, T.; Greasley, P. J.; Wilstermann, M. Scaffold hopping, synthesis and structure-activity relationships of 5, 6-diaryl-pyrazine-2-amide derivatives: A novel series of CB1 receptor antagonists. *Bioorg. Med. Chem.* **2007**, *15*, 4077–4084.

(8) Jennings, A.; Tennant, M. Selection of molecules based on shape and electrostatic similarity: proof of concept of "electroforms. *J. Chem. Inf. Model.* **2007**, *47*, 1829–1838.

(9) Good, A. C. Novel DOCK clique driven 3D similarity database search tools for molecule shape matching and beyond: adding flexibility to the search for ligand kin. *J. Mol. Graphics Modell.* **2007**, *26*, 656–666.

(10) Srinivasan, J.; Castellino, A.; Bradley, E. K.; Eksterowicz, J. E.; Grootenhuis, P. D. J.; Putta, S.; Stanton, R. V. Evaluation of a novel shape-based computational filter for lead evolution: Application to thrombin inhibitors. *J. Med. Chem.* **2002**, *45*, 2494–2500.

(11) Kearsley, S. K.; Smith, G. M. An alternative method for the alignment of molecular structures: Maximizing electrostatic and steric overlap. *Tetrahedron Comput. Method.* **1990**, *3*, 615–633.

(12) Miller, M. D.; Sheridan, R. P.; Kearsley, S. K. SQ: a program for rapidly producing pharmacophorically relevent molecular superpositions. *J. Med. Chem.* **1999**, *42*, 1505–1514.

(13) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.-F.; Cornell, W. D. Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* **2007**, *47*, 1504–1519.

(14) Hawkins, P. C.; Skillman, A. G.; Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **2007**, *50*, 74–82.

(15) Lee, H. S.; Choi, J.; Kufareva, I.; Abagyan, R.; Filikov, A.; Yang, Y.; Yoon, S. Optimization of high throughput virtual screening by combining shape-matching and docking methods. *J. Chem. Inf. Model.* **2008**, *48*, 489–497.

(16) Nicholls, A.; McGaughey, G. B.; Sheridan, R. P.; Good, A. C.; Warren, G.; Mathieu, M.; Muchmore, S. W.; Brown, S. P.; Grant, J. A.; Haigh, J. A. Molecular shape and medicinal chemistry: a perspective. *J. Med. Chem.* **2010**, *53*, 3862–3886.

(17) Grant, J.; Gallardo, M.; Pickup, B. A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *J. Comput. Chem.* **1996**, *17*, 1653–1666.

(18) Sperandio, O.; Andrieu, O.; Miteva, M. A.; Vo, M. Q.; Souaille, M.; Delfaud, F.; Villoutreix, B. O. Med-sumolig: a new ligand-based screening tool for efficient scaffold hopping. *J. Chem. Inf. Model.* **2007**, *47*, 1097–1110.

(19) Ballester, P. J.; Richards, W. G. Ultrafast shape recognition to search compound databases for similar molecular shapes. *J. Comput. Chem.* **2007**, *28*, 1711–1723.

(20) Good, A. C.; Ewing, T. J. A.; Gschwend, D. A.; Kuntz, I. D. New molecular shape descriptors: application in database screening. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 1–12.

(21) Putta, S.; Lemmen, C.; Beroza, P.; Greene, J. A novel shape-feature based approach to virtual library screening. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1230–1240.

(22) Tawa, G. J.; Baber, J. C.; Humblet, C. Computation of 3D queries for ROCS based virtual screens. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 853–868.

(23) Sherman, W.; Beard, H. S.; Farid, R. Use of an induced fit receptor structure in virtual screening. *Chem. Biol. Drug Des.* **2006**, *67*, 83–84.

(24) Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.* **2006**, *49*, 534–553.

(25) Putta, S.; Beroza, P. Shapes of things: computer modeling of molecular shape in drug discovery. *Curr. Top. Med. Chem.* **2007**, *7*, 1514–1524.

(26) Connolly, M. L. Computation of molecular volume. *J. Am. Chem. Soc.* **1985**, *107*, 1118–1124.

(27) Grant, J.; Pickup, B. A Gaussian description of molecular shape. *J. Phys. Chem.* **1995**, *99*, 3503–3510.

(28) Jones, G.; Willett, P.; Glen, R. C. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 532–549.

(29) Richmond, N. J.; Willett, P.; Clark, R. D. Alignment of three-dimensional molecules using an image recognition algorithm. *J. Mol. Graphics Modell.* **2004**, *23*, 199–209.

(30) Krystek, S. R., Jr; Hunt, J. T.; Stein, P. D.; Stouch, T. R. Three-dimensional quantitative structure-activity relationships of sulfonamide endothelin inhibitors. *J. Med. Chem.* **1995**, *38*, 659–668.

(31) Jonker, R.; Volgenant, A. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing* **1987**, *38*, 325–340.

(32) Ferro, D. R.; Hermans, J. A different best rigid-body molecular fit routine. *Acta Crystallogr., Sect. A* **1977**, *33*, 345–347.

(33) *LigPrep v2.4*, Schrödinger, Inc.: Portland, OR, 2010.

(34) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747–750.

(35) Watts, K. S.; Dalal, P.; Murphy, R. B.; Sherman, W.; Friesner, R. A.; Shelley, J. C. ConfGen: A conformational search method for efficient generation of bioactive conformers. *J. Chem. Inf. Model.* **2010**, *50*, 534–546.

(36) Dixon, S.; Smondyrev, A.; Knoll, E.; Rao, S.; Shaw, D.; Friesner, R. PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 647–671.

(37) Salam, N. K.; Nuti, R.; Sherman, W. Novel Method for Generating Structure-Based Pharmacophores Using Energetic Analysis. *J. Chem. Inf. Model.* **2009**, *49*, 2356–2368.

(38) Truchon, J.-F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.

(39) Chen, I. J.; Foloppe, N. Drug-like bioactive structures and conformational coverage with the LigPrep/ConfGen suite: comparison to programs MOE and Catalyst. *J. Chem. Inf. Model.* **2010**, *50*, 822–839.

(40) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical similarity using physiochemical property descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118–127.

(41) Sastry, M.; Lowrie, J. F.; Dixon, S. L.; Sherman, W. Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments. *J. Chem. Inf. Model.* **2010**, *50*, 771–784.