JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

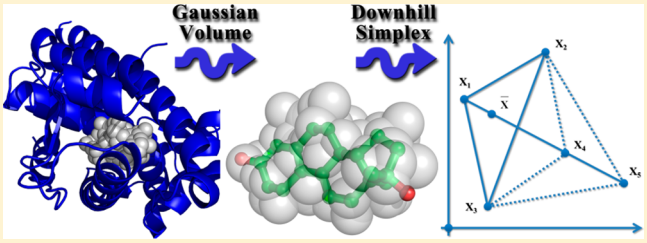
SimG: An Alignment Based Method for Evaluating the Similarity of Small Molecules and Binding Sites

Chaoqian Cai,[†,‡,§] Jiayu Gong,[†,‡,§] Xiaofeng Liu,[‡] Daqi Gao,*[,†] and Honglin Li*[,‡]

[†]School of Information Science and Engineering, and [‡]State Key Laboratory of Bioreactor Engineering, Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, 130 Mei Long Road, Shanghai 200237, China

Supporting Information

ABSTRACT: In this study, a Gaussian volume overlap and chemical feature based molecular similarity metric was devised, and a downhill simplex searching was carried out to evaluate the corresponding similarity. By representing the shapes of both the candidate small molecules and the binding site with chemical features and comparing the corresponding Gaussian volumes overlaps, the active compounds could be identified. These two aspects compose the proposed method named SimG which supports both structure-based and ligand-based strategies. The validity of the proposed method was examined by analyzing the similarity score variation between actives and decoys as well as correlation among distinct reference methods. A retrospective virtual screening test was carried out on DUD data sets, demonstrating that the performance of structure-based shape matching virtual screening in DUD data sets is substantially dependent on some physical properties, especially the solvent-exposure extent of the binding site: The enrichments of targets with less solvent-exposed binding sites generally exceeds that of the one with more solvent-exposed binding sites and even surpasses the corresponding ligand-based virtual screening.



INTRODUCTION

Virtual screening (VS),[1−3] involving the rapid in silico assessment of large libraries of chemical structures in order to identify those most likely to bind to a drug target, can be categorized into ligand-based and structure-based VS methods according to the availability of target protein structure. Ligand-based screening methods are based on the premise so-called "molecular similarity principle"[4] which states that similar molecules will tend to have similar biological properties and thus are more likely to bind to the same target[5−8] while the structure-based methods perform a binding target based screening, taking the structure of the target into consideration. Choosing the appropriate strategy for VS crucially impacts the performance of VS.

Shape complementarity (generally a small molecule and the cavity of a protein target) is pertinent to molecular recognition and usually a prerequisite for protein−ligand binding.[9−15] Currently, typical structure-based VS methods include the popular molecular docking approach[16−20] which calculates the binding affinity of a conformer with specific orientations using a scoring function and guides the molecule to fit into a binding site of given targets. However, recent studies show that it is more challenging to predict the binding affinity accurately.[10,21−25] Alternately, structure-based screening is still applicable to identify novel hits (with weak binding) by means of shape (or chemical feature) matching as long as the shape of the binding site could be properly represented, which is the main concern of this study. Many 3D similarity methods had been developed to date, making it possible to compare the shape of distinct conformers. The Ultrafast Shape Recognition algorithm (USR)[26−28] is a popular method for fast comparison of the 3D shape of molecules, which relies on the relative position of composing atoms to generate a shape descriptor for similarity evaluation. Rapid Overlay of Chemical Structures (ROCS)[29,30] is a widely applied program for molecular similarity evaluation, which employs a gradient based approach to compare the Gaussian volume overlap of two molecules' shapes and key chemotype features. The Spherical Harmonic (SH) method[31−35] is another category of shape comparison that relies on SH projection to compare the surface shape of molecules. In addition, the shape of the binding site can also be detected and represented by some approaches, e.g. the Putative Active Sites with Spheres (PASS)[36] algorithm utilizes a probe filling approach to detect the cavities on protein surface and represent them as a collection of probes. Figure 1 displays a typical case in which the binding site located on the target AR from the Directory of Useful Decoys (DUD)[37,38] data set corresponding to the native ligand was detected and represented by PASS.

In this study, a molecular 3D similarity method named SimG is presented which applies a downhill simplex method to compare the shapes and chemical features of a small molecule and a binding pocket (or ligand) with Gaussian volume overlap

Received: March 5, 2013

   dx.doi.org/10.1021/ci400139j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

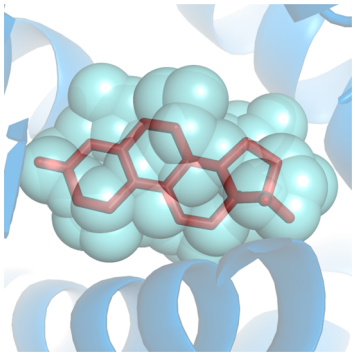


Figure 1. Binding site detected by PASS. Binding site on DUD target AR corresponding to the native ligand was detected by PASS, rendered as a collection of sphere.

in order to identify the active compounds. The basic idea is quite similar to ROCS that the molecular similarity is evaluated by performing a Gaussian volume based alignment, but the proposed method is featured in two aspects. First, both structure-based and ligand-based strategies are supported. When the structure-based strategy is employed, the shape and chemical features of the binding site of a target were extracted and represented using Gaussian function along with small molecules in the screening database, and no information about known ligands was required. While performed with the ligand-based strategy, this method is also capable of evaluating the similarity between two small molecules based on shape and chemical features. Second, a downhill simplex searching algorithm was then carried out to effectively evaluate the shape and chemical feature similarity between database molecules and the query binding site (or ligand) which could be used as a criterion to retrieve active molecules. Compared with existing methods, except the traditional ligand-based similarity methods using the query molecule, SimG also can possess the advantage that it is capable of evaluating the similarity (shape and chemical feature) between a binding site and a small molecule, i.e. binding site based similarity, which will provide another alternative way to discover potential e actives with diverse scaffolds for the given target. This capability is quite useful when no information about the known ligands is available. A retrospective VS test on DUD data set reveals that the performance (within the scope of DUD data sets) of structure-based shape matching VS is substantially dependent on some physical properties, especially the solvent-exposure extent of the corresponding binding site.

MATERIALS AND METHODS

Data Preparation. The DUD database (DUD LIB VS 1.0)[37−40] was chosen as the molecular source to validate and evaluate the performance of the proposed method. Shape representations of the queries were generated by the program PASS: Each DUD target was submitted to PASS using the default parameters (probe radius of 1.8 Å) to enumerate all potential pockets and only the pocket occupied by the native DUD ligand was kept and will be used as the query template to perform VS. These detected pockets represented as pseudo-molecules were then submitted to various molecular similarity evaluation methods (SimG, ROCS, USR, and SH) to carry out similarity based VS. The shape characterization of the binding sites in this study is in some ways similar to the protomol generation step in Surflex.[41] For each pocket shape query, program CASTp[42] was applied to calculate the number and area of the pocket mouth to measure the solvent-exposure of a pocket. These calculated properties are intended for further analysis of the VS results. Four typical binding sites with various degrees of solvent-exposure are depicted in Figure 2, and

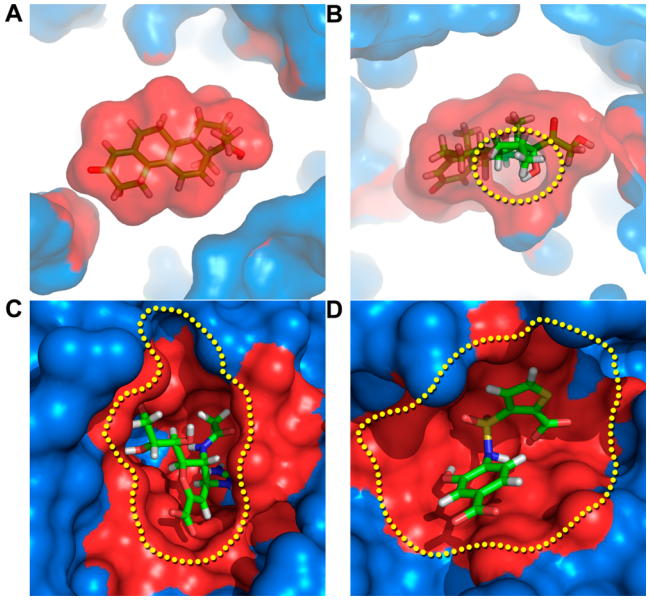


Figure 2. Binding site with various solvent-exposure. Four typical binding sites with various degrees of solvent-exposure extent are rendered as surfaces (highlighted in red) along with binding ligands represented as a stick model. The mouth area of the binding site was struck with a dotted line, and the corresponding mouth areas are (in angstroms squared) (A) AR 0; (B) GR 12.26; (C) NA 73.49; (D) AMPC 476.22.

related analysis will be presented later in this study. After filtering out the targets with inappropriate pocket detected (of which the ligand binding sites were not filled with probes), 24 targets were finally chosen in this study as shown in Table 1. Conformer generation was accomplished using Pipeline Pilot:[43] All actives and decoys were processed by the "3D Conformation" component with a maximum of 50 conformers per molecule. All conformers for the target specific actives and decoys along with the corresponding query template were collected to perform a single run of VS.

Gaussian Volume of Atoms. In order to calculate the Gaussian overlap of atoms, the Gaussian density of a given atom i should be defined as a Gaussian function:[29,44,45]

$$\rho_i(\mathbf{r}_i) = p_i \exp(-\alpha_i \mathbf{r}_i^2) \quad (1)$$

Where, $\alpha_i = \kappa_i/\sigma_i^2$, $\mathbf{r}_i$ is a distance vector originated from the centroid of atom i which is also the Gaussian center, p_i is a parameter determining the maximum value of Gaussian density, σ_i is the radius of atom i, and κ_i is a constant controlling the decay of the density, namely the softness of Gaussian density and whose value will be deduced later. The Gaussian volume of an atom is always equal to the volume of the sphere occupied by the atom, then

$$\int \rho_i(\mathbf{r}_i)\, d\mathbf{r}_i = \frac{4\pi}{3}\sigma^3 \quad (2)$$

Rewrite $\kappa_i = \pi/\lambda_i^{2/3}$ with respect to an artificial variable λ_i, the formula above can be solved as $p_i\lambda_i = 4\pi/3$. Early research

dx.doi.org/10.1021/ci400139j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. Summary of DUD Data Sets Used in This Study[a]

target	PDB code	active no.	decoy no.	cluster no.	pocket volume[b]	mouth no.[b]	mouth area[b]
ACHE	1eve	99	3859	18	1005	1	74.5
ADA	1ndw	23	927	8	514	1	50.42
ALR2	1ah3	26	986	14	1144	2	135.27
AMPC	1xgj	21	786	6	2362	2	476.22
AR	1xq2	68	2848	10	528	0	0
COX1	1q4g	23	910	11	471	2	44.93
ER_AGONIST	1l2i	63	2568	10	614	0	0
ER_ANTAGONIST	3ert	18	1058	8	1476	1	21.34
GART	1c2t	8	155	5	465	1	39.14
GPB	1a8i	52	2135	10	1936	0	0
GR	1m2z	32	2585	9	723	1	12.26
HIVRT	1rt1	34	1494	17	727	3	66.74
HSP90	1uy6	23	975	4	544	3	180.92
INHA	1p44	57	2707	23	1107	3	159.62
MR	2aa2	13	636	2	1134	1	33.33
NA	1a4g	49	1713	7	542	2	73.49
PARP	1efy	31	1350	7	1449	2	193.5
PDE5	1xp0	26	1698	22	1087	1	111.25
PNP	1b8o	25	1036	4	349	1	9.22
RXR	1mvc	18	575	3	1092	1	33.47
SAHH	1a7a	33	1346	2	361	0	0
THROMBIN	1ba8	23	1148	14	739	2	159.46
TK	1kim	22	891	7	608	1	26.68
TRYPSIN	1bju	9	718	7	222	1	28.42

[a]This table lists the target name, PDB code, number of actives, number of decoys, number of clusters, pocket volume, number of mouth, and area of mouth for each DUD target used in this study. [b]These properties are calculated by CASTp.[42]

works[44] have already demonstrated an optimal choice of p_i as $p_i = 2\sqrt{2}$ (and κ_i will also be determined as a constant), and this value will be adopted to implement Gaussian volume calculation throughout this study. On the basis of Gaussian density, the overlapped volume formed by arbitrary atoms from 1 to n can be written as follows according to Gaussian theory:

$$V_{1,2,..,n} = \int \rho_1(\mathbf{r}_1)\rho_2(\mathbf{r}_2)...\rho_n(\mathbf{r}_n)\,\mathrm{d}\mathbf{r} = (\prod_{i=1}^{n} p_i)\exp\left(-\frac{\sum_{i=1,j>1}^{n} \alpha_i\alpha_j R_{ij}^{\ 2}}{\sum_{i=1}^{n}\alpha_i}\right)\left(\frac{\pi}{\sum_{i=1}^{n}\alpha_i}\right)^{3/2} \quad (3)$$

Where, R_{ij} is the distance between the centroid of atoms i and j, $\mathbf{r}$ is a vector representing the global coordinate, and each local coordinate $\mathbf{r}_i$ (i = 1, 2, .., n) can be represented as $\mathbf{r}_i = \mathbf{r} - \mathbf{R}_i$ where $\mathbf{R}_i$ is the centroid of atom i.

Shape Similarity of Molecules. The Tanimoto coefficient (Tc), also known as Jaccard index, is a statistic widely used for comparing the similarity and diversity of two sample sets, which is defined as the size of the intersection divided by the size of the union of the sample sets.[8,46] Given two molecules A and B, their Tanimoto similarity (Tc score) can be written as

$$J_{AB} = \frac{|A \cap B|}{|A \cup B|} = \frac{V_{AB}}{V_A + V_B - V_{AB}} \quad (4)$$

Where, V_A and V_B are the volume of molecules A and B respectively, and V_{AB} denotes the overlapped volume of molecules A and B. A more generalized form of Tc can be written as the Tversky index as follows:

$$T_{AB} = \frac{|A \cap B|}{|A \cap B| + \alpha|A - B| + \beta|B - A|} \quad (5)$$

Where α and β are weight constants. By setting $\alpha = \beta = 1$, the Tversky index is equivalent to the Tc. The overlapped volume can be represented in an integrated form of the density, as follows:

$$V_{AB} = \int P_A P_B\,\mathrm{d}\mathbf{r} = \sum_{i\in A,j\in B} V_{ij} - \sum_{i,j\in A,k\in B} V_{ijk} - \sum_{i\in A,j,k\in B} V_{ijk} + ... \quad (6)$$

Where, P_A and P_B are the Gaussian density of molecules A and B, respectively, $i \in$ A and $j \in$ B denotes the consisting atom of molecules A (with m atoms) and B (with n atoms), respectively. By expanding each individual atom overlap term in eq 6 using eq 3, the overlapped volume of two molecules (V_{AB}) can be calculated analytically. The Gaussian volume of single molecule (V_A and V_B) can be regarded as an overlapped volume with itself and thus can be computed in a similar way: $V_A = V_{AA}$ and $V_B = V_{BB}$. Some recent study[47] employed a simplified measurement scheme also based on overlapped volume but using directly calculated volume as an approximation instead of calculating Gaussian volume.

Chemical Feature Similarity of Molecules. Besides the volume overlap, the chemical features of molecules and binding sites are also taken into consideration in the process of similarity evaluation. There are currently six types of chemical features in total which have been implemented: hydrogen donor/acceptor, positively/negatively charged, and aromatic and hydrophobic type. When a ligand was adopted as a query template, every constituent atom will be assigned one or more chemical features with nonconflict types (e.g., hydrogen donor

dx.doi.org/10.1021/ci400139j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

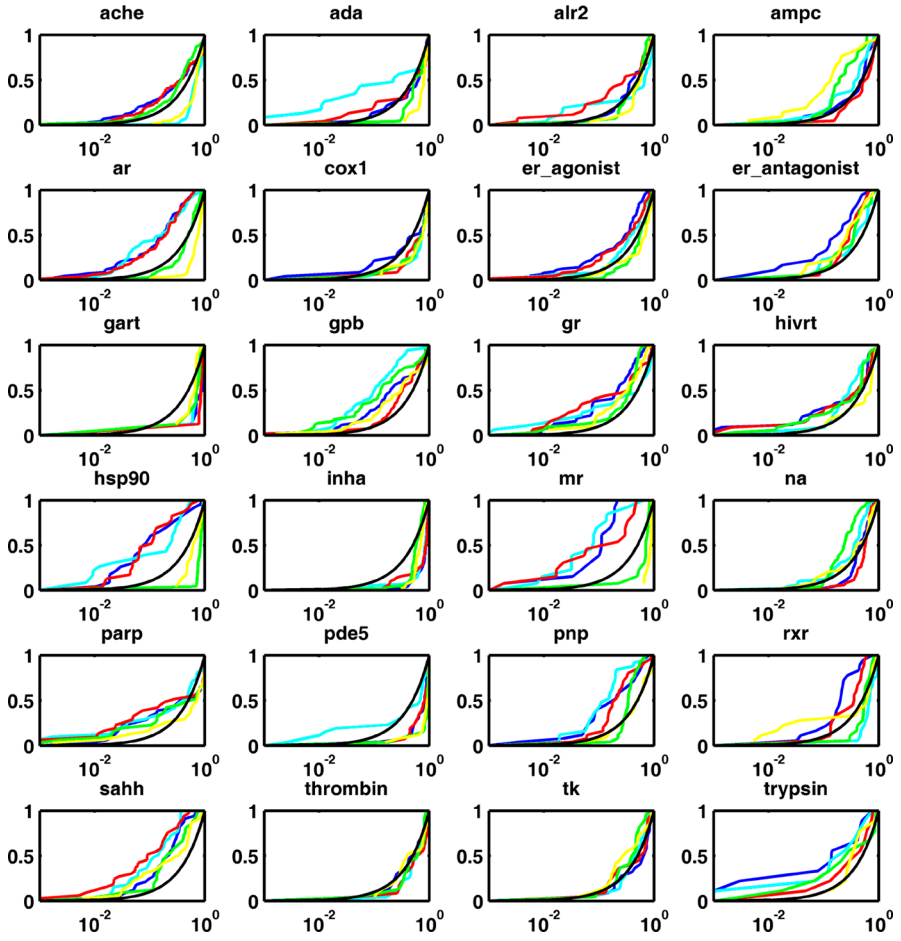


Figure 3. ROC plots for each data set. The *x*-axis corresponds to false positive rate (1 − specificity) while the *y*-axis corresponds to true positive rate (sensitivity). In order to emphasize the early enrichment, the *x*-axis was logarithmically calibrated while the *y*-axis was kept uniformly calibrated. Each method was represented as: blue for SimG (Shape); cyan for SimG (Combo); red for ROCS (Shape); green for USR; yellow for SH; black for random selection.

conflicts with hydrogen acceptor, positively charged type conflicts with negatively charged type) if possible. When a binding site was adopted as a query template, every pocket probe (pseudoatom) will be assigned one or more chemical features if possible even if those features are with conflict types. The chemical features will be utilized to evaluate the molecular alignment in the succeeded optimization process. The chemical feature types for small molecular atoms could be directly deduced from the molecular structure according to some known patterns which are summarized in Supporting Information Figure S1. However, the chemical feature types for pocket probes have to be inferred indirectly from the pocket residues which are defined as any residues within certain distance cutoff (4 Å in this study) away from the corresponding probe, since the pocket probes themselves are merely pseudoatoms containing no chemical information. Depending on the type of residue around a given probe, a complementary chemical feature type is assigned to the probe, and this process is repeated for all residues around. The chemical feature types for all twenty amino acids residues are summarized in Supporting Information Table S1.

Given a specific alignment of a pair of molecules (reference molecule and fit molecule), if atom X from a reference molecule and atom Y from a target molecule are located within some certain distance cutoff (1.4 Å in this study) and both possess a common feature F, then atom X and atom Y are said to be matched with feature F. On the basis of the number of matched atoms, a feature score could be defined to measure the chemical features' consistence of two molecules as follows

$$S_{\text{feature}} = \frac{1}{N}\sum_{i=1}^{N}\frac{n_i^{\text{matched}}}{n_i^{\text{total}}} \tag{7}$$

Where, N is the total number of feature types in a reference molecule, n_i^{matched} is the number of matched atoms with feature i in a reference molecule and n_i^{total} is the total number of atoms with feature i in a reference molecule. By combining shape similarity and chemical feature similarity, a final combo score measuring the similarity between the reference molecule A and fit molecule B could be defined as

$$S = wS_{\text{shape}} + (1 - w)S_{\text{feature}} \tag{8}$$

.Where, S_{shape} is the shape score of molecules A and B calculated using eq 4 while the constant w is a predefined weight controlling the relative contribution of shape score and feature score. In this study, w = 0.5 was chosen for employing a ligand as the query while w = 0.25 was chosen for employing the binding site as the query. A downhill simplex algorithm based optimization was carried out to perform molecular alignment which was described in detail in the Supporting Information.

dx.doi.org/10.1021/ci400139j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 2. AUC Values for Targets[a]

solvent exposure	target	SimG (Shape)	SimG (Combo)	ROCS (Shape)[b]	USR[b]	SH[b]
less exposed	AR	0.79	0.77	0.77	0.50	0.28
	ER_AGONIST	0.68	0.57	0.62	0.44	0.39
	GPB	0.65	0.83	0.58	0.72	0.59
	SAHH	0.76	0.83	0.84	0.70	0.65
	PNP	0.72	0.80	0.72	0.63	0.47
	GR	0.73	0.53	0.65	0.66	0.64
	ER_ANTAGONIST	0.77	0.54	0.67	0.62	0.67
	TK	0.42	0.52	0.49	0.60	0.59
normal	TRYPSIN	0.73	0.70	0.58	0.56	0.56
	MR	0.88	0.87	0.79	0.32	0.16
	RXR	0.75	0.29	0.67	0.44	0.51
	GART	0.19	0.27	0.14	0.30	0.37
	COX1	0.45	0.30	0.35	0.33	0.32
	ADA	0.44	0.57	0.45	0.36	0.24
	HIVRT	0.59	0.61	0.58	0.61	0.44
	NA	0.51	0.63	0.44	0.73	0.55
more exposed	ACHE	0.55	0.29	0.55	0.59	0.32
	PDE5	0.26	0.45	0.25	0.15	0.14
	ALR2	0.50	0.42	0.59	0.51	0.43
	THROMBIN	0.46	0.40	0.36	0.43	0.46
	INHA	0.19	0.21	0.24	0.40	0.32
	HSP90	0.79	0.78	0.82	0.17	0.34
	PARP	0.49	0.57	0.52	0.47	0.38
	AMPC	0.54	0.60	0.42	0.73	0.77
	avg	*0.58*	*0.56*	*0.55*	*0.50*	*0.44*

[a]AUC values achieved by various methods on distinct targets are sorted by the mouth area of the binding site in an ascending order. The last row of the table summarizes the average AUC value. [b]These methods are fully described in other references.[26−30,35]

Performance Evaluation. Evaluation of VS is never a trivial task and usually error-prone due to the absence of robust metrics.[48,49] To visually manifest the performance of VS, the receiver operating characteristic (ROC) curve was employed which is plotted as true positive rate against corresponding false positive rate.[50,51] Because it is not straightforward to compare ROC curves directly, especially for a cross-data set comparison, other statistics designated to quantify the enrichment of true positive hits were also reported. The area under curve (AUC) of ROC[50] is reported as an overall measurement of the performance prior to subsequent detailed measurements. Traditionally, the detailed performance at distinct screening stages is differentiated by enrichment factor (EF) which assesses the improvement of true positive rate over random selection, but EF suffers from a significant dependency on the ratio of actives and decoys under certain conditions.[52,53] Alternatively, as a complement to the global AUC metric, ROC enrichment (ROCE)[39,54] was reported in this study which is defined as the ratio of true positive rate and false positive rate at some specific stage when a particular percentage of decoys is observed. On the other hand, to validate the scaffold-hopping ability of the proposed method, the arithmetic weighted AUC (awAUC) and ROCE (awROCE)[39,51] were reported in parallel, which emphasizes the retrieval of distinct scaffold other than individual active. In addition, analysis concerning the score distribution and correlation was also performed.

RESULTS AND DISCUSSION

Virtual Screening Performance with the Pseudoligands Derived from the Binding Site. In order to examine the ability of the proposed method to retrieve actives using binding site as the query template in a molecular similarity based manner, VS test was performed on DUD data sets. Provided with identical query templates (binding sites extracted by PASS), actives, and target-specific decoys, SimG and other popular molecular shape comparison methods as references are applied to the specific molecule set of each target, to score and rank the candidates. Note that the binding site representation generated by PASS contains only shape information and hence only shape similarity is evaluated by ROCS (Rapid Overlay of Chemical Structures) although ROCS is also capable of evaluating the chemical feature similarity. The general criterion for evaluating the VS performance is to ensure that as many of the top hits as possible are actives.

The ROC curves for each method on each data set were plotted in Figure 3 in different colors to provide a direct visual comparison for the VS performance. The *y*-axis in ROC plot denotes that true positive rate or sensitivity which measures the portion of actives have been correctly identified while the *x*-axis denotes false positive rate or 1 − specificity which measures the cost to achieve the corresponding hit rate, so ROC curve shifting toward the upper left of the coordinate plane generally manifests better performance. The *x*-axis was logarithmically calibrated in order to emphasize the early enrichment while the *y*-axis was kept uniformly calibrated as the recommendation proposed in some previous studies.[51] It can be perceived from the ROC plot that SimG demonstrates an excellent performance compared with other reference methods: It outperformed SH and USR significantly in many cases, and in most cases, its overall performance (Table 2) is very close to that of ROCS; no obvious inferiority in performance was observed. Since it is not quite straightforward to compare ROC curves directly, more detailed and quantitative comparisons will be exhibited in the following text.

dx.doi.org/10.1021/ci400139j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Before looking into the detailed results of VS test in depth, the overall performance is scrutinized in terms of AUC first. Although AUC is a statistic reflecting only the global performance and lacks the capability to differentiate the performance discrepancy at different screening stages especially the early stages, it still gives an outline of the VS performance and could be a complement of other metrics.[54] The AUC values achieved by distinct methods are listed in Table 2 which is intentionally sorted by the mouth area in the ascending order to emphasize the corresponding dependence. The performance for random selection corresponds to an AUC value of 0.5 which is usually referenced as a qualitative level, and any value below this bound is generally considered as a sign of failure. It can be observed clearly from Table 2 that the overall enrichment of structure-based VS test displays a strong dependence on the solvent-exposure extent of binding site. Focused on the shape score scheme of the proposed method, for the first 12 targets corresponding to less solvent-exposed binding sites in the table, 10 cases achieved a relative high performance (AUC > 0.6) and only 2 cases failed (AUC < 0.5), but for the last 12 targets corresponding to more solvent-exposed binding sites, the number of high performance cases drops to 1 while the number of failure cases rises to 6. In another aspect, the first 12 targets gave an average AUC of 0.67 which decreased to 0.48 for the last 12 targets. This dependence on solvent-exposure of binding site can still be perceived when narrowing the range of the top and bottom targets. Similar trends can also be observed from other reference methods. Furthermore, some concrete examples from Figure 2 have revealed this kind of dependence too. The binding site on target AR (Figure 2A) is a completely closed pocket with no exposure of ligand in which case it preserves the accurate shape requirement for a binding ligand. Thus, for target AR it produced a rather high performance for both shape-score-based and combo-score-based schemes (an AUC of 0.79 and 0.77, respectively). On the contrary, the binding site on target AMPC (Figure 2D) is an extremely open gap with a large portion of the ligand exposed to the solution in which case it is rather difficult to reflect the shape occupancy for the binding ligand. Thus, for target AMPC it produced a very poor performance almost no better than random selection. More similar examples could be found, e.g. target GR and NA (Figure 2B and C). In another aspect, the overall VS performance seems also closely related to the volume of binding site adopted as the query: binding sites with small volumes tend to produce high AUCs (e.g., AR, SAHH, PNP, TRYPSIN, HSP90, etc.), while binding sites with large volumes tend to produce relatively low AUCs (e.g., INHA, ALR2, PDE5, AMPC, PARP, etc.). In addition, the proposed method produced an average AUC of 0.58 ± 0.18 which is higher than other reference methods tested in this study (0.55 ± 0.18 for ROCS, 0.50 ± 0.16 for USR, and 0.44 ± 0.16 for SH) and demonstrates the capability of retrieving a large amount of true actives.

VS in practice usually concerns only about a small portion of top hits in the ranking list due to the reason that it is impractical to experimentally test the binding affinity of all candidates.[52] To emphasize the performance at the early stages of the retrieval, the enrichment of actives is measured using the ROCE metric as shown in Figure 4 in which each stacked bar represents an average ROCE at a specific stage. A total of four stages (0.5%, 1%, 2%, and 5%) were chosen as the recommendation proposed in some previous studies.[54] It can be observed clearly that the average ROCE of SimG based on

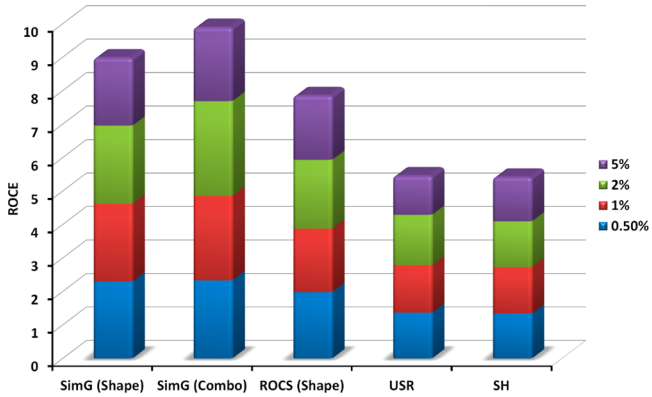


Figure 4. ROCE values achieved by each method. Average ROCE values achieved by each method at different stages are rendered as stacked bars and the height of the bar corresponds to the quantity of ROCE value. The standard deviations for SimG (Shape), SimG-(Combo), ROCS(Shape), USR, and SH are: 3.1, 3.2, 2.8, 2.0, and 1.9 at 0.5% stage; 2.5, 3.0, 2.3, 1.9, and 1.8 at 1% stage; 2.2, 2.5, 2.3, 2.0, and 1.9 at 2% stage; 1.9, 2.4, 2.1, 1.8, and 1.8 at 5% stage.

combo score at all four stages (4.4 at 0.5% stage, 4.2 at 1% stage, 3.9 at 2% stage, and 2.9 at 5% stage) exceeds all other reference methods, indicating a larger portion of true positive top hits and a better discrimination power at the early stages. The combo score scheme also exhibits an appreciable performance gain over the shape score scheme, owing to the combination of chemical features.

More detailed enrichment on each target is listed in Table 3 which is also sorted by the mouth area of binding sites in an ascending order. Quite similar to the performance measured in terms of AUC, the early enrichment listed in Table 3 also discloses a dependence on the solvent-exposure extents of binding sites. Focused on the shape score scheme of the proposed method, for the first 12 targets corresponding to less solvent-exposed binding sites, there are only 4 cases in which no actives were identified at 1% stage while this figure increases to 8 for the last 12 targets corresponding to more solvent-exposed binding sites. The average ROCE achieved by the first 12 targets reaches 5.0 which is much higher compared to an average ROCE of 1.5 achieved by the last 12 targets. A similar trend can be observed at other stages as well as in other methods. This kind of dependence has also been revealed by some concrete examples from Figure 2. The binding site on target GR (Figure 2B) is an almost closed cavity with nearly the whole ligand swallowed by the protein in which case the complementary shape requirement can still be reflected by the binding site. Thus, for target GR, it produced a rather high early enrichment. On the contrary, the relatively more solvent-exposed binding site on target NA (Figure 2C) is a wide and shallow cavity exposing to a large portion of the binding ligand. In such cases, the shape occupancy for the binding ligand cannot be accurately described by the binding site itself, thus it produced a poor early enrichment (failed at all four stages) for target NA. More similar examples could be found, e.g. targets AR and AMPC (Figure 2A and D).

At the 5% stage, both SimG and ROCS performed closely with the proposed method taking a slight advantage. It is still troublesome for USR to retrieve actives effectively at this stage, either in terms of nonzero ROCE count or the average ROCE value compared to other methods. For most of those cases in which the proposed method was unsuccessful to retrieve any

Table 3. ROCE Values Achieved on Targets[a]

		SimG (Shape)				SimG (Combo)				ROCS (Shape)[b]				USR[b]				SH[b]			
solvent exposure	target	0.50%	1%	2%	5%	0.50%	1%	2%	5%	0.50%	1%	2%	5%	0.50%	1%	2%	5%	0.50%	1%	2%	5%
less exposed	AR	11.0	7.2	5.8	4.4	2.8	2.9	3.6	6.7	5.5	5.7	4.4	4.1	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0
	ER_AGONIST	3.1	7.8	5.4	3.5	6.2	3.1	3.1	1.3	3.1	3.1	3.1	2.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6
	GPB	3.6	1.9	3.8	2.7	0.0	5.6	5.8	6.9	3.6	1.9	1.0	1.1	7.1	13.1	8.6	5.0	3.6	1.9	1.9	2.7
	SAHH	0.0	0.0	1.5	1.8	0.0	0.0	4.4	4.8	5.9	8.8	7.3	6.1	0.0	2.9	2.9	1.8	0.0	2.9	1.5	4.2
	PNP	7.1	3.9	2.0	1.6	0.0	0.0	5.9	2.4	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8
	GR	0.0	6.2	6.2	3.7	17.8	12.4	7.7	5.0	0.0	6.2	9.3	6.9	0.0	3.1	4.6	1.9	5.9	3.1	3.1	1.2
	ER_ANTAGONIST	9.4	5.1	8.1	3.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.3
	TK	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9
normal	TRYPSIN	37.1	21.2	10.6	4.4	18.5	10.6	5.3	2.2	0.0	0.0	0.0	2.2	0.0	0.0	5.3	2.2	0.0	0.0	0.0	0.0
	MR	13.7	6.9	3.7	4.6	0.0	13.7	7.5	7.6	13.7	6.9	11.2	6.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	RXR	0.0	0.0	0.0	2.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.3	10.3	5.5
	GART	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	COX1	7.4	4.1	2.2	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	ADA	0.0	0.0	0.0	0.0	21.4	11.9	12.6	7.6	0.0	0.0	2.1	2.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	HIVRT	15.6	8.3	4.3	4.1	5.2	2.8	1.4	1.2	15.6	8.3	4.3	2.9	5.2	2.8	4.3	2.9	0.0	0.0	0.0	0.6
	NA	0.0	0.0	0.0	0.0	0.0	0.0	1.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	1.0	1.2
more exposed	ACHE	0.0	3.0	4.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	2.0	3.9	2.0	1.5	1.2	0.0	0.0	0.0	0.0
	PDE5	0.0	0.0	0.0	0.0	7.5	7.5	7.5	3.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	ALR2	0.0	0.0	0.0	0.8	0.0	0.0	5.5	3.8	14.0	7.0	3.7	3.0	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0
	THROMBIN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	0.0	0.0	0.0	0.9	0.0	0.0	0.0	0.0
	INHA	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	HSP90	0.0	0.0	8.3	5.2	7.5	20.7	12.4	6.9	0.0	0.0	6.2	6.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	PARP	0.0	3.2	3.2	4.5	19.0	9.5	6.3	5.1	12.6	6.3	7.9	6.4	12.6	6.3	4.7	3.9	6.3	3.2	3.2	2.6
	AMPC	0.0	0.0	0.0	1.9	0.0	0.0	2.3	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	8.7	4.4	7.0	3.8
	avg	*2.3*	*2.3*	*2.3*	*2.0*	*2.3*	*2.5*	*2.8*	*2.2*	*2.0*	*1.9*	*2.1*	*1.9*	*1.4*	*1.4*	*1.5*	*1.2*	*1.3*	*1.4*	*1.4*	*1.3*

[a]ROCE values are evaluated at four different stages (0.5%, 1%, 2%, and 5%) with respect to each method. All targets are ordered by the mouth area in an ascending manner, and the last row of the table counts the average ROCE values over the above targets. [b]These methods are fully described in other references.[26−30,35]

dx.doi.org/10.1021/ci400139j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

actives in the early stages, the other reference methods also failed, e.g. AMPC, GART, INHA, NA, and PDE5. It is understandable to expect such a relatively high failure rate in early stages due to the fact that using a negative image of the binding pocket as the query template is not guaranteed to preserve the exact molecular shape of the corresponding ligand. In addition, some representative binding site examples (ER_ANTAGONIST and TRYPSIN) in which the performance of SimG and ROCS varied significantly are illustrated in Figure S5 in the Supporting Information.

Scaffold Hopping Potential. Scaffold hopping potential is another crucial aspect related to VS performance which reflects the capability to retrieve novel scaffolds or chemotypes.[48,55] The DUD database has provided the scaffold based clusters ready to be utilized in scaffold hopping validations[39] which will be adopted in this study. Similar to the metric of AUC and ROCE but concentrating on scaffolds instead of individual active, awAUC and awROCE were employed to manifest the scaffold hopping ability as shown in Figure 5 and Supporting Information Table S2. The awROCE was evaluated separately at 0.5%, 1%, 2%, and 5% stages in accordance with reports previously shown in this study. Over all targets, the proposed method achieved an average awAUC of 0.56 ± 0.20 that is the highest among reference methods, followed by ROCS (0.54 ± 0.20), USR (0.46 ± 0.15), and, then, SH (0.43 ± 0.17) in turn. The proposed method also achieved the highest awROCE at 5% stage although ROCS took the lead in some earlier stages with a slight disparity. Generally, USR and SH method performed rather closely in the aspect of awROCE with SH method possessing a tiny advantage in all four specific stages.

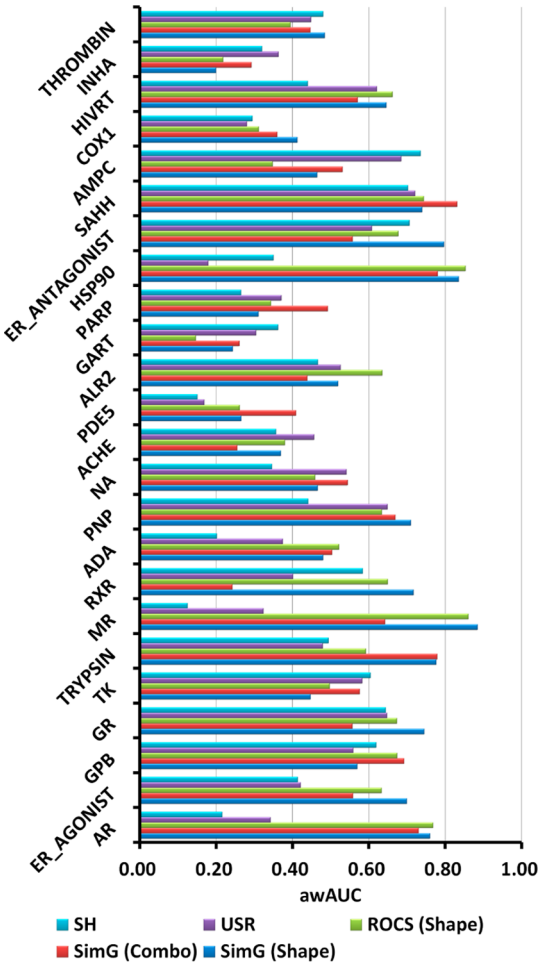


Figure 5. awAUC values on each data set. awAUC values achieved by various methods for distinct targets are rendered as groups of horizontal bars with different colors. The *x*-axis quantifies the awAUC value while the *y*-axis is labeled with target names.

Correlation of Similarity Ranks Obtained from Various Methods. Besides assessing the VS performance of various molecular similarity comparison methods, it is significant to explore the internal relationship of similarity ranks obtained from these distinct methods. However, the quantities of individual similarity scores calculated from different methods cannot be compared directly due to the fact that they represent similarity measurements under metrics of different kinds. But it is still nonetheless possible and reasonable to sketch out the cross-methods rank correlation in a statistical manner which was widely adopted by many studies.[8,15,56−60] Figure 6 manifests the similarity rank correlation between these molecular similarity comparison methods as scatter plots in which the ordinate denotes a rank computed by the experimental method in this study while the abscissa denotes a rank acquired from various other control methods. All sample points in this figure are obtained from the VS test described previously, and the upper three plots correspond to SimG shape rank while the middle three plots correspond to SimG combo rank (evaluated through a combination of shape score and feature score). In the ultimate ideal case which indicates that both the experimental method and the control method could measure the molecular similarity unambiguously, then there will be a bijection between experimental ranks and control ranks, meaning that each distinct experimental rank can be mapped to a nonrepeated control rank and vice versa. Under this circumstance, the strict one to one correspondence will exhibit a functionlike image in the scatter plot since the experimental rank and control rank can be regarded as functions of each other reciprocally.

In the SimG shape rank versus ROCS shape rank case (Figure 6A), most sample points are distributed along the main diagonal in a long and narrow region, exhibiting an obvious linear trend and thus suggesting a close coincidence of these two methods. The clearly observed linear relationship between these ranks corroborates the validity of the proposed ranking method from another aspect besides the VS performance, since ROCS is doubtlessly a long validated and widely applied molecular shape comparison method. On the other hand, the introduction of feature score significantly affects the linear trend with ROCS shape rank as shown in Figure 6D, which demonstrates that a great contribution has been made from the chemical feature. In both the SimG versus USR and SH cases (panels B, C, E, and F in Figure 6), most of the sample points are scattered in a broad area deviating from the main diagonal, and no evident function like correlation can be perceived from the plot. To maintain the integrity of the correlation analysis, the correlation between ROCS, USR, and SH ranks are also plotted in Figure 6 (panels G, H, and I).

Comparison of Using Binding Site and Ligand as the Query Template. Both the binding site and the native ligand from the complex structure preserve shape information that can be availed to perform VS; thus, both of them can be chosen as the query templates which correspond to the structure-based strategy and ligand-based strategy respectively. In this section, it focused on using the native ligand as the query template. In order to compare the ability to retrieve active compounds

dx.doi.org/10.1021/ci400139j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

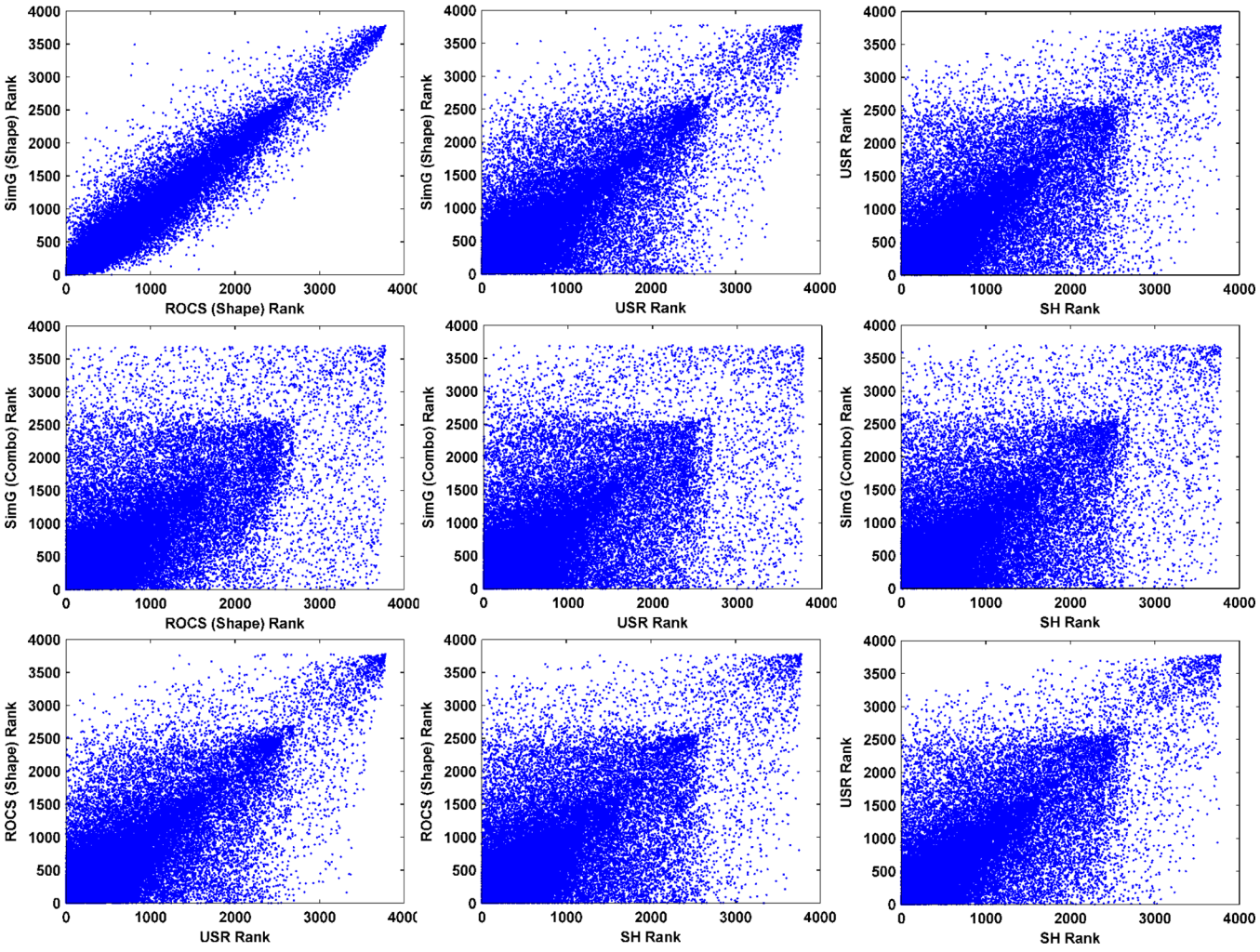


Figure 6. Correlations of similarity ranks produced by various molecular similarity comparison methods: (A) SimG Shape versus ROCS Shape; (B) SimG Shape versus USR; (C) SimG Shape versus SH; (D) SimG Combo versus ROCS Shape; (E) SimG Combo versus USR; (F) SimG Combo versus SH; (G) ROCS Shape versus USR; (H) ROCS Shape versus SH; (I) USR versus SH.

between the structure-based and the ligand-based strategies, another VS test was performed on the same conformer collection but used the single native ligand as query template extracted from a bound complex structure and the corresponding results measured in AUC are listed in Table 4 which are also sorted by the mouth area in ascending order. When comparing the results from Tables 2 and 4, it can be observed that the structure-based strategy is comparable to the ligand-based strategy and even outperformed it on targets with less solvent-exposed binding sites (top items in the table) although the structure-based strategy gave a relatively lower average performance over all targets. Focused on the proposed method, for the shape score scheme there are four cases out of the first eight targets (with an average AUC of 0.69) in which the structure-based strategy outperformed the ligand-based strategy and produced a better-than-random result, while this figure drops to one for the last eight targets (with an average AUC of 0.47). Looking closely into the examples presented in Figure 2, a consistent result could be perceived: The performance of structure-based strategy outperformed corresponding ligand-based strategy on targets AR and GR which possess relatively closed binding sites; meanwhile, it failed dramatically on targets NA and AMPC that possess relatively greater solvent-exposed binding sites. The results obtained by combo score scheme or ROCS method also gave a similar trend, exhibiting a strong dependence on the solvent-exposure extent of binding sites. This particular dependence may be explained in the aspect of shape matching. On one hand, it is much more difficult to represent the exact shape of an intensely solvent-exposed binding site. On the other hand, it is much more possible for a ligand with great shape variation to bind to an intensely solvent-exposed pocket, e.g. partial structure of the ligand may protrude from the binding site or the ligand only occupy a relatively small portion of the binding site. More concrete examples will be given in the following text.

In order to further explore the performance discrepancy caused by the shape divergence of the binding site and native ligand, the binding site (rendered as gray spheres) detected by PASS and the native ligand (rendered as stick model) from a complex structure in some typical cases were overlaid in Figure 7. As shown in Figure 7A and B, the binding site strategy outperformed native ligand strategy significantly in MR and ER_ANTAGONIST cases. For the target MR, the native ligand is a four-ring structure with higher rigidity and fully occupies the binding site, and in this case, the shape of the binding site may be more suitable to reflect the shape complement requirement for a binding ligand. For the target ER_ANTAGONIST, the native ligand is a relatively flexible structure with

dx.doi.org/10.1021/ci400139j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 4. AUC Values for Targets by Using a Native Ligand As the Query Template[a]

solvent exposure	target	SimG (Shape)	SimG (Combo)	ROCS[b] (Shape)	ROCS[b] (Combo)	USR[b]	SH[b]
less exposed	AR	0.76	0.67	0.75	0.73	0.56	0.52
	ER_AGONIST	0.50	0.81	0.50	0.82	0.51	0.40
	GPB	0.79	0.89	0.80	0.93	0.73	0.56
	SAHH	0.88	0.91	0.91	0.94	0.76	0.76
	PNP	0.78	0.80	0.77	0.89	0.77	0.65
	GR	0.55	0.53	0.50	0.63	0.46	0.50
	ER_ANTAGONIST	0.75	0.89	0.75	0.86	0.52	0.49
	TK	0.76	0.82	0.77	0.84	0.67	0.69
normal	TRYPSIN	0.55	0.56	0.53	0.44	0.52	0.60
	MR	0.86	0.81	0.91	0.92	0.71	0.58
	RXR	0.91	0.84	0.92	0.99	0.83	0.65
	GART	0.65	0.75	0.60	0.61	0.71	0.64
	COX1	0.51	0.58	0.52	0.54	0.37	0.50
	ADA	0.52	0.67	0.51	0.69	0.40	0.48
	HIVRT	0.65	0.64	0.70	0.68	0.71	0.47
	NA	0.86	0.88	0.86	0.92	0.79	0.58
more exposed	ACHE	0.69	0.70	0.71	0.76	0.66	0.61
	PDE5	0.61	0.61	0.60	0.63	0.58	0.45
	ALR2	0.31	0.37	0.30	0.55	0.42	0.31
	THROMBIN	0.54	0.46	0.42	0.48	0.42	0.28
	INHA	0.61	0.53	0.60	0.57	0.64	0.41
	HSP90	0.75	0.81	0.72	0.90	0.65	0.54
	PARP	0.45	0.65	0.50	0.57	0.43	0.37
	AMPC	0.86	0.82	0.85	0.78	0.80	0.75
	avg	*0.67*	*0.71*	*0.67*	*0.73*	*0.61*	*0.53*

[a]Applying a native ligand as the query template, AUC values achieved by various methods on distinct targets are sorted by the mouth area of the binding site in an ascending order. The last row of the table summarizes the average AUC value. [b]These methods are fully described in other references.[26−30,35]

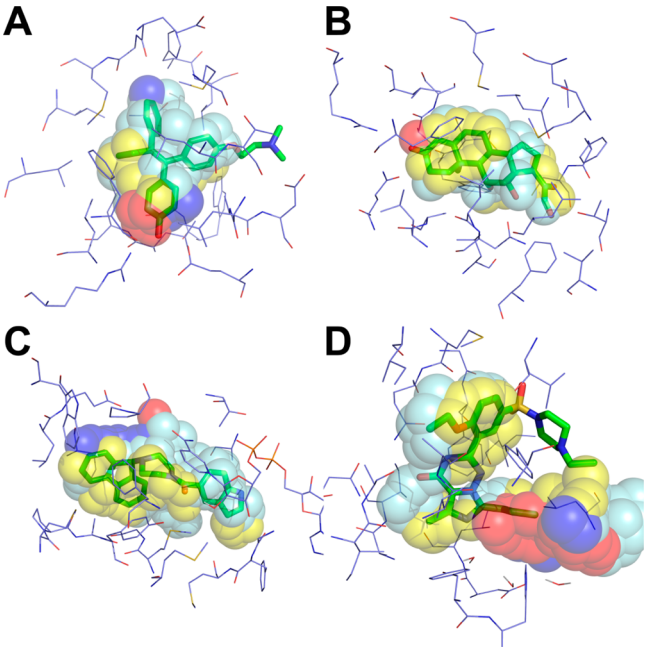


Figure 7. Divergence in shape between binding site and native ligand. The shape of binding site detected by PASS was rendered as colored spheres [(red) hydrogen bond acceptor, (blue) hydrogen bond donor, (yellow) aromatic] and the native ligand was rendered as a stick model. These four cases are the following: (A) ER_ANTAGONIST; (B) MR; (C) INHA; (D) PDE5.

a dimethylamino-ethoxy group protruding from the binding pocket, and in this case, the shape of binding site is a partial match of the native ligand and thus may also be able to identify the binding ligands with complementary shape. As shown in Figure 7B and C, the binding site strategy failed dramatically compared with the native ligand strategy in the INHA and PDE5 cases. For the target INHA, the native ligand is completely embedded into a cavity, but a large portion of the binding site is still not occupied, leaving a vacant space, and in this case, using the entire shape of binding site as a shape matching criterion may result in a large number of false positives especially for those large size decoys since any decoy with a larger overlapped volume will produce a higher similarity score. For the target PDE5, the native ligand occupied only one of the two subpockets with a 1-ethyl-4-sulfonyl-pyrazine group exposed outside the binding site, and in this case, the entire shape of binding site may also be inappropriate to be taken as a query template to perform shape matching. It has been demonstrated in the above examples that a binding site with full shape occupancy with the entire or partial native ligand generally preserves sufficient shape information to identify the actives from decoys while a binding site occupied partially by the native ligand (especially a ligand with relatively small volume) is more apt to give a high false positive rate.

Comparison with Docking Methods. Many VS methods concern some aspects other than shape or chemical feature matching, e.g. docking-based binding affinity prediction. In order to make a comparison between the proposed method and those docking-based methods, some performance statistics are extracted from related literature. Some previous studies[61] made an intense evaluation of many popular docking methods (DOCK,[62] Surflex,[41] FRED,[30] FlexX,[63] and ICM[64]) on 11 DUD data sets but using enrichment rate at 1% stage as a

dx.doi.org/10.1021/ci400139j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 5. Enrichment Rate at 1% Stage Compared with Some Popular Docking Methods[a]

target	SimG (Pocket Shape)	SimG (Pocket Combo)	SimG (Ligand Shape)	SimG (Ligand Combo)	Surflex[b]	FRED[b]	FlexX[b]	ICM[b]	DOCK[b]
ADA	0.00%	13.04%	4.35%	8.70%	7.69%	5.13%	5.13%	0.00%	15.00%
ER_ANTAGONIST	5.56%	0.00%	22.22%	27.78%	17.95%	12.82%	15.38%	17.95%	10.00%
HIVRT	8.82%	2.94%	14.71%	14.71%	18.60%	18.60%	0.00%	11.63%	0.05%
NA	0.00%	0.00%	24.49%	22.45%	22.45%	0.00%	4.08%	34.69%	10.00%
THROMBIN	0.00%	0.00%	0.00%	4.35%	5.56%	1.39%	6.94%	5.56%	5.00%
TK	0.00%	0.00%	18.18%	9.09%	0.00%	18.18%	0.00%	0.00%	0.00%
TRYPSIN	22.22%	11.11%	33.33%	11.11%	18.37%	2.04%	0.00%	2.04%	0.00%

[a]The proposed method using binding site and native ligand as the query template are labeled as "SimG Pocket" and "SimG Ligand". [b]The corresponding data are derived from other references.[61]

measurement, and the performance statistics on 7 data sets coinciding with this study were listed in Table 5. It can be observed that the performance on target ER_ANTAGONIST (relatively more solvent-exposed) for both shape-based methods and docking-based methods is very close to each other while the shape-based methods produced a much higher enrichment on targets TK and TRYPSIN both possessing a relatively more closed binding site. On the other hand, the docking-based methods take a tiny advantage over shape-based methods on target THROMBIN. It seems that the docking-based methods are less affected by the solvent-exposure extent of corresponding binding site.

CONCLUSIONS

In this study, a Gaussian volume overlap and chemical feature based molecular similarity metric was devised, and a downhill simplex searching was carried out to effectively evaluate the corresponding similarity. By representing not only the candidate small molecules but also the binding site as Gaussian volume with chemical features and comparing the corresponding volumes, the active compounds could be identified. These two aspects compose the proposed method named SimG which supports both structure-based and ligand-based strategies. The validity of the proposed method was examined by analyzing the similarity score variation between actives and decoys as well as correlation among distinct reference methods. A retrospective VS test was performed on DUD data sets, demonstrating that the performance of structure-based shape matching VS on DUD data sets is substantially dependent on some physical properties of corresponding binding site, especially the solvent-exposure extent of the binding site: The enrichment of the VS test on less solvent-exposed binding sites generally exceeds that on more solvent-exposed binding sites and even surpasses the corresponding ligand-based VS.

ASSOCIATED CONTENT

Supporting Information

The downhill simplex algorithm described in detail. Other discussion includes: the score variation and the computation time required by the molecular alignment. The figure for chemical feature definition and the table for awROCE summary are also provided. This material is available free of charge via the Internet at http://pubs.acs.org. The executable program of SimG is available at http://lilab.ecust.edu.cn/home/resource.html.

AUTHOR INFORMATION

Corresponding Author

*Phone: +86-21-64250213. Fax: +86-21-64250213. E-mail: dqgao@ecust.edu.cn (D.G.); hlli@ecust.edu.cn (H.L.).

Author Contributions

§C.C. and J.G.: These authors contributed equally to this paper.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the Fundamental Research Funds for the Central Universities, the National Natural Science Foundation of China (grants 21173076, 81102375, 81230090, 81222046 and 81230076), the Special Fund for Major State Basic Research Project (grant 2009CB918501), the Shanghai Committee of Science and Technology (grants 11DZ2260600 and 12401900801), and the 863 Hi-Tech Program of China (grant 2012AA020308). H.L. is also sponsored by Program for New Century Excellent Talents in University (grant NCET-10-0378) and Shanghai Rising-Star Tracking Program (grant 13QH1401100).

REFERENCES

(1) Gorse, A. D. Diversity in medicinal chemistry space. *Curr. Top. Med. Chem.* **2006**, *6*, 3−18.

(2) Ripphausen, P.; Nisius, B.; Peltason, L.; Bajorath, J. Quo vadis, virtual screening? A comprehensive survey of prospective applications. *J. Med. Chem.* **2010**, *53*, 8461−8467.

(3) Reddy, A. S.; Pati, S. P.; Kumar, P. P.; Pradeep, H. N.; Sastry, G. N. Virtual screening in drug discovery—a computational perspective. *Curr. Protein. Pept. Sci.* **2007**, *8*, 329−351.

(4) Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C. K.; Glick, M.; Davies, J. W. How similar are similarity searching methods? A principal component analysis of molecular descriptor space. *J. Chem. Inf. Model.* **2009**, *49*, 108−119.

(5) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.

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

(7) Bender, A.; Glen, R. C. Molecular similarity: A key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204−3218.

(8) Nikolova, N.; Jaworska, J. Approaches to measure chemical similarity - A review. *Qsar. Comb. Sci.* **2004**, *22*, 1006−1026.

(9) Gohlke, H.; Klebe, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem., Int. Ed.* **2002**, *41*, 2645−2676.

(10) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **2007**, *50*, 74−82.
(11) Meyer, M.; Wilson, P.; Schomburg, D. Hydrogen bonding and molecular surface shape complementarity as a basis for protein docking. *J. Mol. Biol.* **1996**, *264*, 199−210.
(12) Amovilli, C. Shape and similarity: Two aspects of molecular recognition. *J. Mol. Struct.−THEOCHEM* **1991**, *227*, 1−9.
(13) Liebeschuetz, J.; Hennemann, J.; Olsson, T.; Groom, C. R. The good, the bad and the twisted: A survey of ligand geometry in protein crystal structures. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 169−183.
(14) Tasi, G.; Palinko, I.; Molnar, A.; Hannus, I. Molecular shape, dimensions, and shape selective catalysis. *J. Mol. Struct.−THEOCHEM* **2003**, *666*, 69−77.
(15) 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.; Nevins, N.; Jain, A. N.; Kelley, B. Molecular shape and medicinal chemistry: A perspective. *J. Med. Chem.* **2010**, *53*, 3862−3886.
(16) Lengauer, T.; Rarey, M. Computational methods for biomolecular docking. *Curr. Opin. Struct. Biol.* **1996**, *6*, 402−406.
(17) Shoichet, B. K.; Bodian, D. L.; Kuntz, I. D. Molecular docking using shape descriptors. *J. Comput. Chem.* **1992**, *13*, 380−397.
(18) Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. Distributed automated docking of flexible ligands to proteins: Parallel applications of AutoDock 2.4. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 293−304.
(19) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat. Rev. Drug Discovery* **2004**, *3*, 935−949.
(20) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved protein-ligand docking using GOLD. *Proteins* **2003**, *52*, 609−623.
(21) 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.
(22) Perez-Nueno, V. I.; Ritchie, D. W.; Rabal, O.; Pascual, R.; Borrell, J. I.; Teixido, J. Comparison of ligand-based and receptor-based virtual screening of HIV entry inhibitors for the CXCR4 and CCR5 receptors using 3D ligand shape matching and ligand-receptor docking. *J. Chem. Inf. Model.* **2008**, *48*, 509−533.
(23) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912−5931.
(24) Englebienne, P.; Moitessier, N. Docking ligands into flexible and solvated macromolecules. 4. Are popular scoring functions accurate for this class of proteins? *J. Chem. Inf. Model.* **2009**, *49*, 1568−1580.
(25) Stouch, T. R. The errors of our ways: Taking account of error in computer-aided drug design to build confidence intervals for our next 25 years. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 125−134.
(26) Ballester, P. J.; Richards, W. G. Ultrafast shape recognition to search compound databases for similar molecular shapes. *J. Comput. Chem.* **2007**, *28*, 1711−1723.
(27) Ballester, P. J.; Finn, P. W.; Richards, W. G. Ultrafast shape recognition: Evaluating a new ligand-based virtual screening technology. *J. Mol. Graphics Modell.* **2009**, *27*, 836−845.
(28) Ballester, P. J. Ultrafast shape recognition: method and applications. *Future Med. Chem.* **2011**, *3*, 65−78.
(29) 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.
(30) *ROCS*, version 2.2; OpenEye Scientific Software: Santa Fe, NM, 2006.
(31) Mak, L.; Grandison, S.; Morris, R. J. An extension of spherical harmonics to region-based rotationally invariant descriptors for molecular shape description and comparison. *J. Mol. Graphics Modell.* **2008**, *26*, 1035−1045.
(32) Morris, R. J.; Najmanovich, R. J.; Kahraman, A.; Thornton, J. M. Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics* **2005**, *21*, 2347−2355.
(33) DiMaio, F. P.; Soni, A. B.; Phillips, G. N.; Shavlik, J. W. Spherical-harmonic decomposition for molecular recognition in electron-density maps. *Int. J. Data Min. Bioin.* **2009**, *3*, 205−227.
(34) Mavridis, L.; Hudson, B. D.; Ritchie, D. W. Toward high throughput 3D virtual screening using spherical harmonic surface representations. *J. Chem. Inf. Model.* **2007**, *47*, 1787−1796.
(35) Cai, C.; Gong, J.; Liu, X.; Jiang, H.; Gao, D.; Li, H. A novel, customizable and optimizable parameter method using spherical harmonics for molecular shape similarity comparisons. *J. Mol. Model.* **2012**, *18*, 1597−1610.
(36) Brady, G. P.; Stouten, P. F. W. Fast prediction and visualization of protein binding pockets with PASS. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 383−401.
(37) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.
(38) A Directory of Useful Decoys. http://dud.docking.org/ (accessed Jun 13, 2010).
(39) Jahn, A.; Hinselmann, G.; Fechner, N.; Zell, A. Optimal assignment methods for ligand-based virtual screening. *J. Cheminf.* **2013**, *1*, No. Article 14, http://www.jcheminf.com/content/1/1/14, (accessed Apr 30, 2013).
(40) Good, A. C.; Oprea, T. I. Optimization of CAMD techniques 3. Virtual screening enrichment studies: A help or hindrance in tool selection? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 169−178.
(41) Jain, A. N. Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46*, 499−511.
(42) Dundas, J.; Ouyang, Z.; Tseng, J.; Binkowski, A.; Turpaz, Y.; Liang, J. CASTp: Computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res.* **2006**, *34*, W116−W118.
(43) *Pipeline Pilot*, version 7.5; Accelrys: San Diego, CA, 2008.
(44) Grant, J. A.; Pickup, B. T. A Gaussian description of molecular shape. *J. Phys. Chem.* **1995**, *99*, 3503−3510.
(45) Good, A. C.; Hodgkin, E. E.; Richards, W. G. The utilisation of Gaussian functions for the rapid evaluation of molecular similarity. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 188−191.
(46) Maldonado, A. G.; Doucet, J. P.; Petitjean, M.; Fan, B. T. Molecular similarity and diversity in chemoinformatics: From theory to applications. *Mol. Diversity* **2006**, *10*, 39−79.
(47) Sastry, G. M.; Dixon, S. L.; Sherman, W. Rapid shape-based ligand alignment and virtual screening method based on atom/feature-pair similarities and volume overlap scoring. *J. Chem. Inf. Model.* **2011**, *51*, 2455−2466.
(48) Geppert, H.; Vogt, M.; Bajorath, J. Current trends in ligand-based virtual screening: Molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205−216.
(49) Nicholls, A. What do we know and when do we know it? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 239−255.
(50) Fawcett, T. An introduction to ROC analysis. *Pattern Recog. Lett.* **2006**, *27*, 861−874.
(51) Clark, R. D.; Webster-Clark, D. J. Managing bias in ROC curves. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 141−146.
(52) 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.
(53) Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection–what can we learn from earlier mistakes? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 213−228.

(54) Jain, A. N.; Nicholls, A. Recommendations for evaluation of computational methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 133−139.
(55) Mackey, M. D.; Melville, J. L. Better than random? The chemotype enrichment problem. *J. Chem. Inf. Model.* **2009**, *49*, 1154−1162.
(56) Marques, J. M. C.; Llanio-Trujillo, J. L.; Abreu, P. E.; Pereira, F. B. How different are two chemical structures? *J. Chem. Inf. Model.* **2010**, *50*, 2129−2140.
(57) Swamidass, S. J.; Baldi, P. Mathematical correction for fingerprint similarity measures to improve chemical retrieval. *J. Chem. Inf. Model.* **2007**, *47*, 952−964.
(58) Kotani, T.; Higashiura, K. Rapid evaluation of molecular shape similarity index using pairwise calculation of the nearest atomic distances. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 58−63.
(59) Haigh, J. A.; Pickup, B. T.; Grant, J. A.; Nicholls, A. Small molecule shape-fingerprints. *J. Chem. Inf. Model.* **2005**, *45*, 673−684.
(60) Perez-Nueno, V. I.; Rabal, O.; Borrell, J. I.; Teixido, J. APIF: A new interaction fingerprint based on atom pairs and its application to virtual screening. *J. Chem. Inf. Model.* **2009**, *49*, 1245−1260.
(61) Giganti, D.; Guillemain, H.; Spadoni, J. L.; Nilges, M.; Zagury, J. F.; Montes, M. Comparative evaluation of 3D virtual ligand screening methods: Impact of the molecular alignment on enrichment. *J. Chem. Inf. Model.* **2010**, *50*, 992−1004.
(62) Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 411−428.
(63) Rarey, M.; Kramer, B.; Lengauer, T. Time-efficient docking of flexible ligands into active sites of proteins. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1995**, *3*, 300−308.
(64) Schapira, M.; Totrov, M.; Abagyan, R. Prediction of the binding energy for small molecules, peptides and proteins. *J. Mol. Recognit.* **1999**, *12*, 177−190.

dx.doi.org/10.1021/ci400139j | *J. Chem. Inf. Model.* XXXX, XXX, XXX−XXX