JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

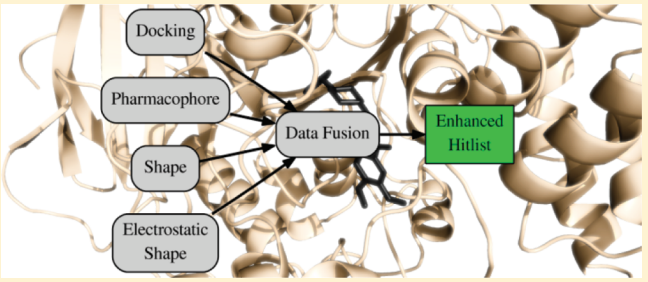
Virtual Screening Data Fusion Using Both Structure- and Ligand-Based Methods

Fredrik Svensson, Anders Karlén, and Christian Sköld*

Organic Pharmaceutical Chemistry, Department of Medicinal Chemistry, BMC, Uppsala University, P.O. Box 574, SE-751 23 Uppsala Sweden

Supporting Information

ABSTRACT: Virtual screening is widely applied in drug discovery, and significant effort has been put into improving current methods. In this study, we have evaluated the performance of compound ranking in virtual screening using five different data fusion algorithms on a total of 16 data sets. The data were generated by docking, pharmacophore search, shape similarity, and electrostatic similarity, spanning both structure- and ligand-based methods. The algorithms used for data fusion were sum rank, rank vote, sum score, Pareto ranking, and parallel selection. None of the fusion methods require any prior knowledge or input other than the results from the single methods and, thus, are readily applicable. The results show that compound ranking using data fusion improves the performance and consistency of virtual screening compared to the single methods alone. The best performing data fusion algorithm was parallel selection, but both rank voting and Pareto ranking also have good performance.



INTRODUCTION

Virtual screening (VS), using both structure- and ligand-based techniques, is today an integral part of the drug discovery process.[1,2] Structure-based methods require knowledge about the target's 3D struture, while ligand-based approaches use information from at least one known ligand. The method of choice often depends on the amount of information that is available. When possible, both structure- and ligand-based methods can be used together in a VS campaign. Often, this is done in a sequential fashion filtering down the number of compounds in each step.[3−7]

Many methods have been applied for VS.[1,8] One common method is molecular docking, i.e., fitting ligand molecules into the active site of a protein in order to predict the binding affinity. Other strategies involve screening compound databases for matches against pharmacophore sites deemed important for biological activity and similarity screening based on the principle that compounds with similar molecular shapes have similar properties.[9]

An important challenge in VS is to create accurate scoring and ranking functions that can distinguish between active and inactive molecules. For docking, the problems involved in creating scoring functions have recently been highlighted.[10,11] Data fusion combines results from several sources aiming to produce a result superior to any of the individual sources. When a data fusion algorithm calculates an entity's new score considering its score in all of the different methods, it is often referred to as a consensus algorithm. Using several different scoring functions followed by fusion of the scores to a consensus score is an approach that is now commonly used in docking.[12] It is also possible to improve the results from VS by using the consensus of several methods employed in parallel.[7,13−20] This has been shown in idealized computer experiments to, at least in part, depend on the statistical factor that the mean value of a number of measurements tends to be closer to the actual value then a single measurement.[21] Baber et al. suggest that increased performance using data fusion is due to VS methods having a higher correlation when ranking active compounds than when ranking inactive ones.[16] Thus, fusing several methods will tend to aggregate active compounds more than inactive ones. Other studies, however, raise the question of whether the results from some of the commonly used methods are complementary and, therefore, best applied using parallel selection methods.[13,22] Even though data fusion is an established tool for VS, there are still relatively few reports on data fusion from both ligand- and structure-based sources.[7,13,20]

Using well-defined public benchmarking data sets should facilitate cross study comparison of results and make a study more transparent. Evaluation of VS methodology depends heavily on the data sets used, and one widely used benchmarking data set is the directory of useful decoys (DUD) compiled by Huang et al.[23] The aim of the DUD data set was to avoid bias created by trivial differences between active and inactive compounds and is compiled of decoys that resemble the active compounds in physical properties but that are topologically dissimilar. However, ligand-based methods

Received: October 12, 2011
Published: December 12, 2011

   dx.doi.org/10.1021/ci2004835 | J. Chem. Inf. Model. 2012, 52, 225−232

can easily account for these differences.[24] The subset of DUD proposed by Cheerseright et al. attempts to tackle this problem by taking into account how many different scaffolds exist among the active compounds for each target.[25]

The aim of this study was to employ several common structure- and ligand-based VS methods on a large set of general and representative targets, fuse the results, and evaluate whether this can increase the enrichment compared to the single methods alone. In addition, we also wanted to identify the most suitable data fusion algorithm. We were primarily interested in evaluating this for VS projects early in drug discovery, and therefore we set as a prerequisite that none of the methods should require more information than one crystallized protein−ligand complex.

METHODS

Data Set Preparation. We have used two data set sources, referred to as data set source 1 (DS1) and data set source 2 (DS2). DS1 was previously compiled and published by Jacobson and Karlén[26] and later modified by Mutas et al.[27] It contains six protein targets with a number of known actives taken from the literature as well as 10 000 diverse decoy compounds drawn from a database of commercially available compounds, which were filtered according to Lipinski's rule of five[28] prior to selection. DS2 is a subset of DUD consisting of 10 DUD data sets but with the full collection of decoys from all 40 DUD data sets. Good and Opera filtered the DUD data set for lead likeness and then clustered the results on the basis of reduced graph representation.[29] Cheeseright et al. suggest using a subset of 13 targets with more than 15 clusters.[25] Since these have the greatest difference between actives they should prove more suitable for ligand-based screening methods, and these targets are therefore used in this study. Of these 13 targets, two were excluded due to problems with broken or missing ligands in the crystal structure (CDK2, VEGFr2), and one (PDGFrb) was excluded since it is a homology model. The exclusion of these targets has previously been argued by Kalliokoski et al.[30] All targets used in this study and their respective number of actives are listed in Table 1. The targets fXa and COX-2 overlap between DS1 and DS2. The protein−ligand complexes for fXa are different in DS1 and DS2, but the protein complexes for COX-2 have the same origin. However, for both targets, different sets of active and inactive compounds are used in DS1 and DS2, thus still providing different challenges for the methods. We elected to use the entire DUD database of decoys for our VS against DS2 in order to reflect a real screening scenario where a large number of diverse compounds would be considered. In total 95 171 decoy compounds were included.

The DUD files, release 2, were downloaded from dud.docking.org (March 8, 2011). All structures were prepared using LigPrep[31] and Epik,[32] generating all possible tautomeric, enantiomeric, and protonation states, as well as calculating metal binding states. We have chosen not to distinguish between different enantiomeric forms of the compounds since we assume that they are tested as racemates. Only the best scoring form for each compound was considered.

All protein X-ray structures were downloaded from the Brookhaven Protein Data Bank[33] and prepared using Protein Preparation Wizard.[34] During the preparation, all water molecules were deleted, hydrogen atoms added, the complex minimized, and metal binding states calculated. For proteins containing several identical subunits with a binding site in each, only one subunit was kept. All other settings were kept at the default setting.

Table 1. Protein targets

PDB code	abbreviation	protein	number of active compounds
		DS1	
1H00	CDK2	cyclin-dependent kinase 2	98
6COX	COX-2	cyclooxygenase 2	310
1L2I	ERα	estrogen receptor α	151
1IQE	fXa	factor Xa	155
1CIZ	MMP3	matrix metalloprotease 3	114
1F8D	NA	neuraminidase	50
		DS2	
1O86	ACE	angiotensin-converting enzyme	49
1EVE	AChE	acetylcholinesterase	105
1CX2	COX-2	cyclooxygenase 2	348
1M17	EGFr	epidermal growth factor receptor	444
1F0R	fXa	factor Xa	142
1RT1	HIVRT	HIV reverse transcriptase	40
1P44	InhA	enoyl ACP reductase	85
1KV2	P38	P38 mitogen activated protein kinase	256
1XP0	PDE5	phosphodiesterase 5	51
2SRC	SRC	tyrosine kinase	155

VS Methods. Four methods spanning different principles of VS were applied in this study: Docking using Glide,[35−37] pharmacophore search using Phase,[38] shape similarity using ROCS,[39] and electrostatic similarity using EON.[40] Since the aim was to investigate the effects of different data fusion methods and not the performance of the individual software, the settings for each individual program were not optimized, but rather the methods were set up to reflect standard settings.

For docking, the SP mode in Glide was used, and all settings were left at the default except for adding Epik state penalties to the docking score. The receptor grids were generated using the grid generation in Glide, centered around the crystallized ligand using default settings.

For the ligand-based methods, the crystallized ligand was used in the tautomeric form suggested after preparation by Protein Preparation Wizard.

To generate the pharmacophore features used by Phase, we used the energetic pharmacophore method presented by Salam et al.[41] This method for pharmacophore generations was chosen since it does not require prior knowledge of more than one active compound and thus can be more readily applied at the start of a screening project. The cocrystallized ligands were docked using Glide XP refine docking, and the descriptors generated were used with the E-Pharmacophores script in Maestro to identify energetically favorable interactions with the protein. The locations of these interactions were then used as pharmacophore sites. The "true" pharmacophore, as presented by Gund,[42] would consist of all features that contribute to the biological response of a compound, and it should be noted that the use of the pharmacophore concept herein is not necessarily the true pharmacophore but rather a collection of structural features providing favorable protein−ligand interactions. Since Glide XP descriptors do not include an explicit term for interactions with metals, a pharmacophore site was added manually to the ligand in MMP3 in order to capture this interaction. Phase was used with default settings to generate

dx.doi.org/10.1021/ci2004835 | J. Chem. Inf. Model. 2012, 52, 225−232

Table 2. Data Fusion Algorithms Applied in This Study

algorithm	description
sum rank	This adds together the ranks from the different VS methods' rank lists.
rank vote	Each screening method votes for its 250 highest ranked compounds. The ranking is primarily based on the number of votes each compound has received (between 4 and 0) and secondarily on the compounds' sum score.
sum score	The relative score of each compound in each method is calculated by dividing all of the scores by the best score any compound acquired from that method. The calculated scores for one compound from each of the four methods are then summed.
pareto ranking	Pareto ranking ranks a compound on the basis of how many other compounds are better in all screening methods. Ties are broken using sum rank.
parallel selection	Compounds are selected from the top ranked compounds from each method in turn until the desired number of compounds is reached. If a compound that would be selected has already been selected before, the next compound from that method is chosen instead.

Table 3. EF in the Top 1% of the Data Sets[a]

data set	EF Max[b]	Glide	Phase	ROCS	EON	sum ranks	rank vote	sum scores	Pareto rank	parallel
					DS1					
CDK2	100.0	**57.1**	14.3	15.3	6.1	17.3	22.4	21.4	40.8	41.8
COX-2	33.2	29.4	**33.2**	**33.2**	32.0	**33.2**	**33.2**	**33.2**	**33.2**	32.0
ERα	67.2	33.9	26.6	44.6	12.6	40.6	43.3	**45.2**	33.3	40.6
fXa	65.5	**49.4**	10.3	9.0	12.8	15.4	14.8	14.1	32.7	25.7
MMP3	88.7	**65.0**	24.6	47.4	58.0	48.3	60.6	61.5	58.8	57.1
NA	100.5	64.3	52.2	84.4	82.4	90.4	94.4	**98.5**	88.4	90.4
					DS2					
ACE	100.0	30.6	42.9	59.2	63.3	61.2	61.2	**65.3**	59.2	59.2
AChE	100.0	3.8	2.9	**69.5**	68.6	16.2	65.7	61.9	54.3	64.7
COX-2	100.0	52.6	52.0	72.1	35.9	69.6	74.7	**75.9**	74.2	70.1
EGFr	100.0	5.0	70.5	**76.4**	25.7	55.6	74.6	72.3	57.2	75.5
fXa	100.0	**54.2**	1.1	4.9	1.4	2.1	39.4	7.7	21.1	39.4
HIVRT	100.0	10.0	20.0	17.5	17.5	20.0	**22.5**	20.0	**22.5**	**22.5**
InhA	100.0	31.8	42.3	**49.4**	28.2	36.5	42.3	35.3	43.5	42.3
P38	100.0	0.4	**12.1**	7.8	6.6	9.0	11.3	10.5	11.3	11.3
PDE5	100.0	2.0	17.7	17.7	9.8	11.8	19.6	17.7	**21.6**	19.6
SRC	100.0	**8.4**	0.6	0.0	0.0	0.0	5.2	0.7	3.9	5.2

[a]Bold numbers indicate the highest enrichment against that target. [b]Maximum enrichment factor obtainable in this fraction.

Table 4. EF in the Top 10% of the Data Sets[a]

data set	EF Max[b]	Glide	Phase	ROCS	EON	sum ranks	rank vote	sum scores	Pareto rank	parallel
					DS1					
CDK2	10.0	**8.6**	2.4	2.5	1.9	4.8	**8.6**	7.8	8.4	8.3
COX-2	10.0	9.8	7.1	9.8	9.8	9.9	9.9	9.9	9.9	**10.0**
ERα	10.0	6.1	6.4	6.7	3.6	6.2	6.2	6.0	**6.8**	6.5
fXa	10.0	8.1	2.1	2.5	6.5	8.2	8.3	8.6	**8.7**	8.1
MMP3	9.5	**9.5**	4.5	6.4	7.3	8.1	9.1	9.1	9.0	9.1
NA	10.0	9.4	8.4	**10.0**	9.2	**10.0**	**10.0**	**10.0**	**10.0**	**10.0**
					DS2					
ACE	10.0	6.1	6.9	7.1	**9.6**	8.0	8.2	7.8	7.8	7.6
AChE	10.0	5.1	1.0	7.7	**8.1**	6.5	7.8	7.8	7.7	7.9
COX-2	10.0	8.6	7.0	8.7	6.8	8.6	8.7	8.6	**8.9**	**8.9**
EGFr	10.0	4.7	9.0	9.1	5.2	8.3	8.7	8.6	**9.5**	**9.5**
fXa	10.0	**8.2**	0.9	1.3	0.2	0.8	4.3	3.0	5.7	7.0
HIVRT	10.0	5.0	3.5	2.8	3.3	**5.3**	5.0	5.0	**5.3**	4.3
InhA	10.0	4.1	4.9	5.3	3.4	5.3	5.1	5.1	5.3	**5.4**
P38	10.0	1.6	1.4	1.6	2.1	2.8	**2.9**	**2.9**	2.7	2.7
PDE5	10.0	0.4	**3.5**	2.6	2.0	2.4	2.9	2.9	3.1	**3.5**
SRC	10.0	**2.1**	0.7	0.5	0.6	0.2	0.8	0.5	1.2	1.4

[a]Bold numbers indicate the highest enrichment against that target. [b]Maximum enrichment factor obtainable in this fraction.

conformers of the compounds. For screening, partial matches down to three pharmacophore features were allowed using the default 2 Å tolerance setting.

ROCS requires that the input structures contain all conformers to be used in the calculations. Therefore, OMEGA[43] was used to generate conformers for further calculations. In a study aimed at reproducing bioactive conformations of crystallized ligands, Boström et al.[44] proposed a number of settings for use in OMEGA. These settings were used in this study, and a maximum of 1000 conformers were generated for each molecule. ROCS was set to generate output files for EON, which was subsequently used. Each target's

dx.doi.org/10.1021/ci2004835 | J. Chem. Inf. Model. 2012, 52, 225−232

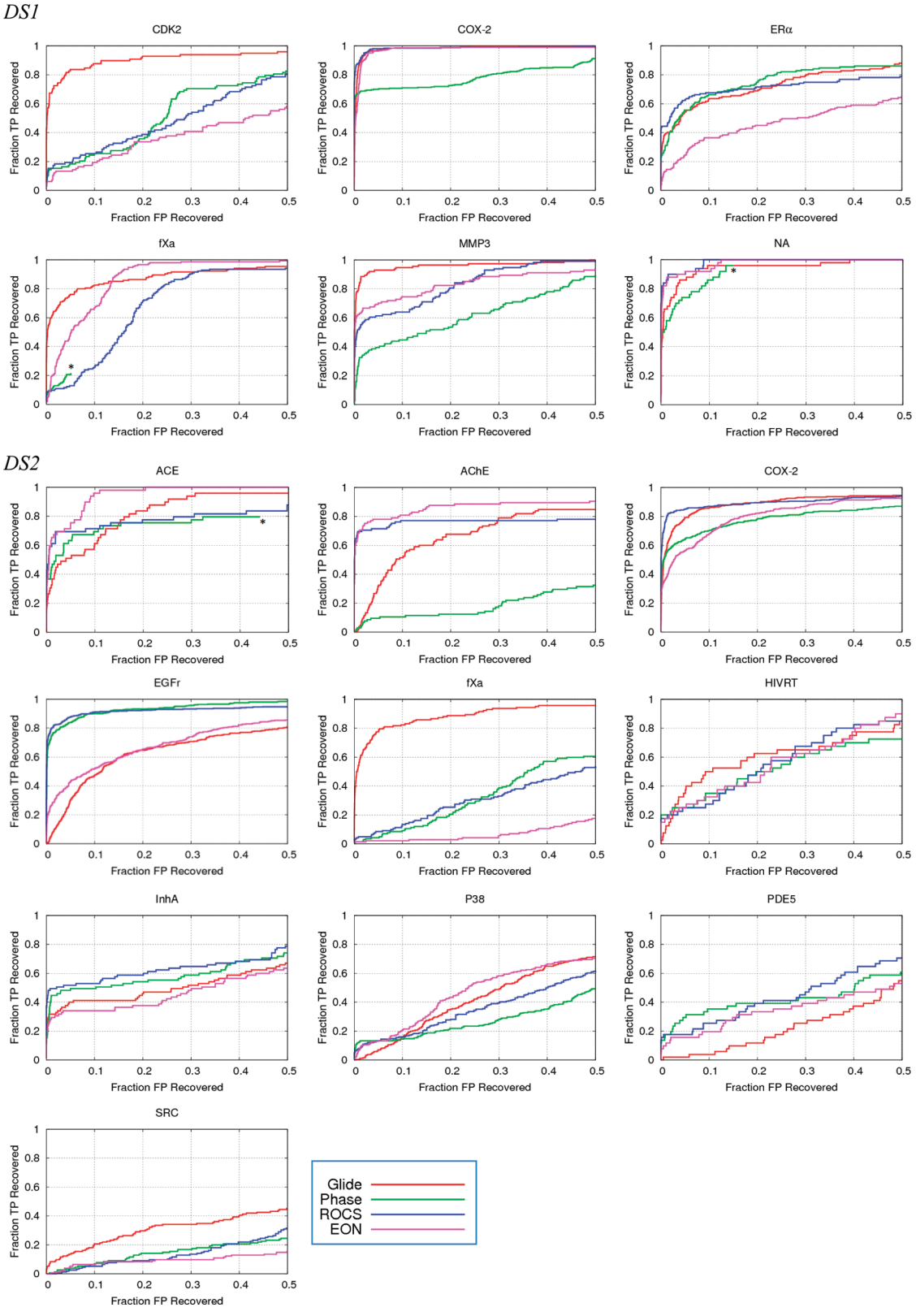


Figure 1. ROC curves for the four VS screening methods. Incomplete lines exist for Phase because not enough compounds could fit the pharmacophore and are marked with an asterisk (*).

respective crystallized ligand was used as a query for both ROCS and EON.

Data fusion. Five different data fusion algorithms were used to generate data fusion ranks, namely, sum rank, rank vote, sum score, Pareto ranking, and parallel selection. The different algorithms are shortly explained in Table 2. These algorithms were chosen for this study since none of them require any training set and thus can be used in a VS scenario without prior

dx.doi.org/10.1021/ci2004835 | J. Chem. Inf. Model. 2012, 52, 225−232

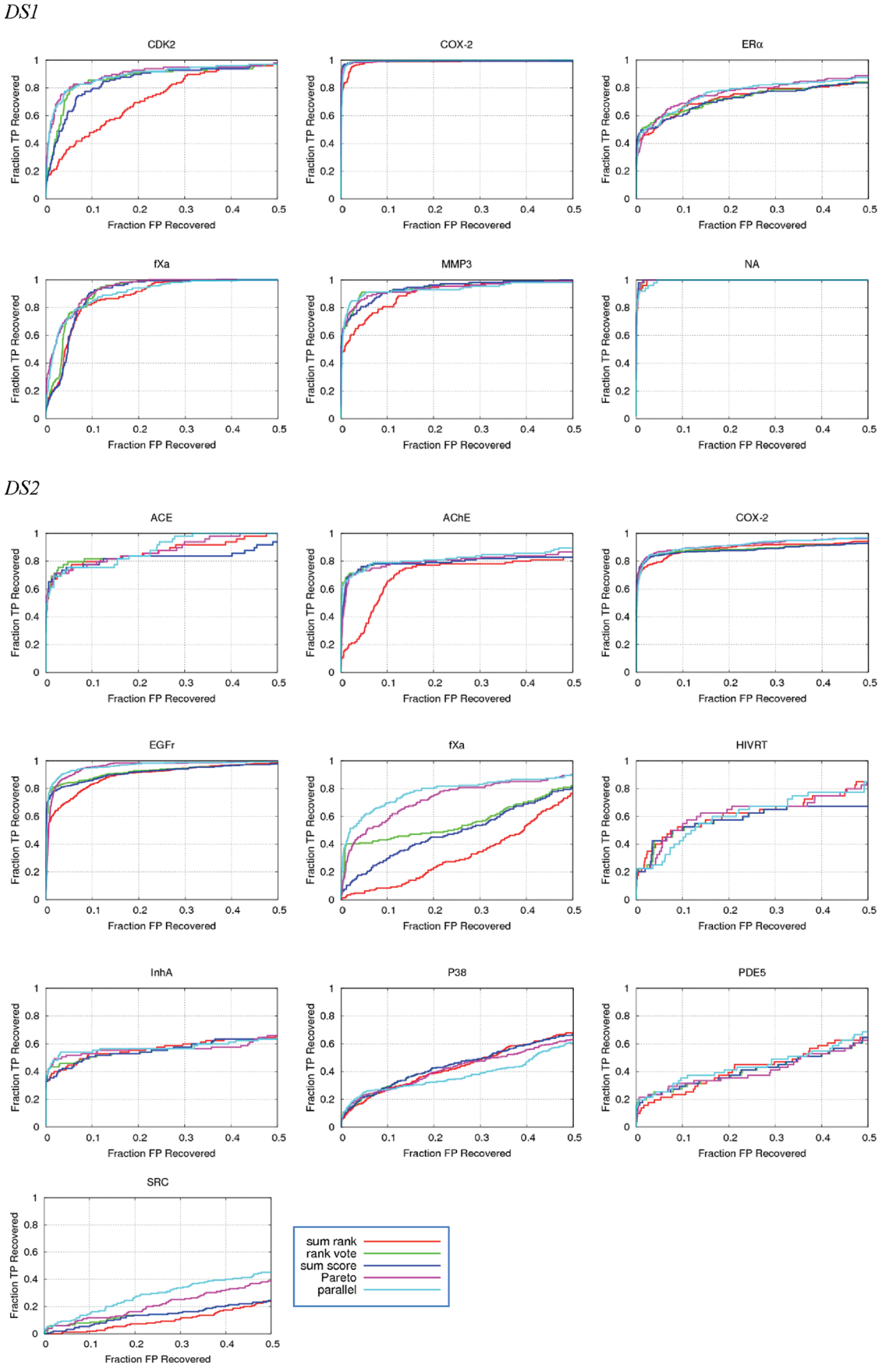


Figure 2. ROC curves for the data fusion algorithms.

dx.doi.org/10.1021/ci2004835 | J. Chem. Inf. Model. 2012, 52, 225−232

Table 5. Average Relative EF in the Top Ranked 100 Compounds, 1%, and 10% of the Data Set[a]

	all data sets			DS1			DS2		
	100 comp.	1%	10%	100 comp.	1%	10%	100 comp.	1%	10%
Glide	0.34	0.44	0.66	**0.68**	**0.68**	0.87	0.13	0.20	0.46
Phase	0.33	0.34	0.46	0.41	0.42	0.52	0.27	0.26	0.39
ROCS	0.43	0.46	0.55	0.56	0.55	0.64	**0.36**	0.37	0.47
EON	0.33	0.37	0.53	0.48	0.48	0.65	0.24	0.26	0.41
sum rank	0.38	0.43	0.64	0.58	0.58	0.79	0.26	0.28	0.48
rank vote	**0.45**	0.52	0.71	0.62	0.62	0.88	0.34	**0.42**	0.54
sum score	0.43	0.50	0.69	0.63	0.63	0.86	0.31	0.37	0.52
Pareto rank	0.40	0.51	**0.73**	0.66	0.66	**0.89**	0.25	0.37	0.57
parallel sel.	0.44	**0.53**	**0.73**	0.65	0.65	0.87	0.32	0.41	**0.58**

[a]Bold numbers indicate the highest average relative EF in the considered fraction of the data set after screening.

knowledge of active compounds. Other studies have reported the use of these or similar methods.[13,16,17,20,21]

Rank voting presents a special challenge since the number of votes given by each method has to be determined. We have chosen to use 250 votes per individual VS method, generating a moderate amount of compounds with votes. Since the rank vote method only assigns 250 votes per VS method and sum score is used for ranking compounds with the same number of votes, the ranking at higher percentages of the database will be the same as for sum score.

Enrichment Factor. To determine the success of the different VS methods and data fusion algorithms, the metric enrichment factor (EF) calculated using eq 1 was used. EF is the relative enrichment of truly active compounds (true positives) in the fraction of database predicted to be active.

$$\mathrm{EF} = \frac{\mathrm{tp}/(\mathrm{tp} + \mathrm{fp})}{A/T} \quad (1)$$

In eq 1, tp is the number of true positives, fp is the number of false positives, A is the total number of active compounds, and T is the total number of compounds in the data set.

Graphical presentations of the results were made using receiver operator characteristics (ROC) curves. A ROC curve plots the true positive rate against the false positive rate.

RESULTS AND DISCUSSION

The results from the VS are presented in Tables 3 and 4. These tables show the EF for each method and for all data sets in the top 1% and 10% of the data sets (results for other fractions are provided in the Supporting Information).

The results show that for all data sets there is no single method that consistently would be the best choice, but rather, there are large differences between the data sets. The results also depend of how large the fraction is of the data sets that are considered after screening. Interestingly, each method recovers the most actives at least once. Comparing the EFs for the data fusion methods with the individual VS methods, there are several data sets where data fusion algorithms outperform any single VS method at both percentages, which shows that data fusion can have more than an averaging effect.

When the data set sources are compared, it appears that some data sets in DS1 do not pose sufficient challenges to the VS methods, with two methods reaching the maximum enrichment against COX-2 at 1% of the database and with very high enrichments against NA at the top 10% of the data sets.

The ROC curves for the VS methods are plotted in Figure 1. The ROC curves further highlight that the performance of the individual VS methods varies between data sets and that there is often a notable variation between different methods within the same target, which makes it difficult to predict what method will perform satisfactorily in a specific case. Furthermore, as seen when comparing Tables 3 and 4, there are many cases where the best performing VS method shifts depending on the size of the top fraction of the data sets considered, which is seen as line crossings in the plots. The poor performance of all methods in the SRC data set is notable. This has also been noted by Diller and Li and was argued to depend on a rather closed kinase structure because of the cocrystallization with an ATP analogue.[45]

The e-pharmacophore method in particular has a high variation between the targets. This can be explained by how the pharmacophore points are selected. Since only interactions that can exist between the cocrystallized ligand and the protein are considered, this method requires that the active compounds in the data sets bind in a similar manner and with the same or a subset of at least three pharmacophore features.

The ROC curves for the data fusion methods are plotted in Figure 2. As in the case for the VS methods, it is difficult to single out a data fusion algorithm that gives the highest EF for all data sets. However, the difference in performance between the data fusion methods is lower compared to the VS methods (Figure 1). Sum rank has the worst performance among the data fusion algorithms in this study. The difference from the similar method sum score can be explained by the large number of molecules in the data sets, which makes sum rank an ineffective tool since a large number of compounds can have very similar scores but still will be assigned quite different ranks. Thus, sum score should provide a more fair result in this case. However, both sum scores and sum rank in general depend on all included VS methods having high performance since a low score for an active compound in one method will penalize that compound with a rather low rank after data fusion. Rank voting, Pareto ranking, and parallel selection are more robust regarding such penalizing effects since they instead focus on only rewarding high ranked hits in any of the individual VS methods.

In order to investigate a general trend, the average relative EF across all targets was calculated. The relative EF is derived by dividing the calculated enrichment factors by the maximum possible EF. The results for the top 100 compounds, the top 1% of the database, and the top 10% of the data sets are presented in Table 5 (results for other fractions are provided in the Supporting Information). When the results of DS1 are compared with those of DS2, it is confirmed that DS2 provides

a more difficult challenge for the individual screening methods but also that there is more to be gained from data fusion in this case, presenting improved results in both the top 1% and the top 10% of the database. Notably, when all 16 data sets are considered, data fusion methods recover active compounds better than any of the individual methods. Even though the performance of all of the data fusion methods is quite similar, some differences can be observed. Among the 100 top ranked compounds for each method, rank voting had the best performance, followed by parallel selection and sum score. When considering the top 1% of the data sets, parallel selection recovered the most actives, which was also the case at the top 10% of the data sets, although Pareto ranking showed the same performance at this fraction of the data sets. However, parallel selection displayed consistently high performance irrespective of the different fractions of the data sets considered after screening (see also Supporting Information). In summary, the results show that parallel selection is a robust and high performing method for recovering actives, which supports previously reported results that have been obtained using fewer targets and smaller data sets[22] or data fusing employing docking and 2D shape methods.[13,22]

To further analyze the consistency of the methods, the number of times each method was among the top three methods was calculated and is presented in Table 6 (an expanded table is provided in the Supporting Information). Rank vote, Pareto rank, and parallel selection have a high occurrence in the top three, showing that they have a consistent performance. In the top 100 compounds and 1% of the data sets, rank voting was among the three methods that gave the highest enrichment in 12 and 14 data sets, respectively, out of 16. However, as previously observed, parallel selection shows good performance in all top fractions of the data sets investigated.

Table 6. The Number of Times Each Method Was among the Top Three Methods in Different Top Fractions of the Data Sets[a]

	100 comp.	1%	3%	5%	10%	15%	20%	50%
Glide	5	5	6	5	5	5	6	10
Phase	4	3	2	1	1	3	2	1
ROCS	10	5	4	4	4	2	3	5
EON	1	2	1	1	2	3	4	5
sum rank	3	3	2	2	5	5	6	8
rank vote	12	14	9	11	10	8	7	5
sum score	9	6	5	6	7	8	6	5
Pareto rank	5	8	13	10	12	14	13	12
parallel sel.	11	10	11	13	11	13	12	13

[a]In cases of equal EF for a top three rank, all of those methods were given the same rank.

Even though the results favor data fusion, a limitation of multiple-method VS is the obvious higher demand for computational resources. It is also difficult to determine how many and what individual VS methods should be included in a VS campaign employing data fusion. However, in the present study, it is clear that, on average, data fusion recovers more actives than any of the four individual VS methods studied. Although this may mainly originate from the data set dependent high variability in the performance of each individual VS method, data fusion gave higher EF compared to any of the individual VS methods used in several data sets. Although rank voting and Pareto ranking are close in performance, parallel selection is the most robust and high performing method for recovering actives.

CONCLUSIONS

In summary, we have performed VS using 16 different data sets, including 14 different targets. We have used four VS methods employing both ligand- and structure-based methods. Subsequently, five data fusion algorithms were applied, and the results were compared with the results generated by the individual methods.

The results show that data fusion is a valuable tool in VS. Data fusion reduces dependency on single VS method performance as well as having the potential to outperform the best single method used. We show that, on average, over all 16 data sets investigated herein, data fusion methods recover more active compounds than any of the individual methods. Among the applied data fusion algorithms, parallel selection has the best average performance in recovering actives in the data sets, but both rank voting and Pareto ranking also have good performances.

ASSOCIATED CONTENT

Supporting Information

EFs for all data sets in the top ranked 100 compounds, 1%, 3%, 5%, 10%, 15%, 20%, and 50% of the data set. Examples of calculated data fusion ranks. E-pharmacophores generated. Flags used for OMEGA. ROC curves with all methods plotted in the same plot for all data sets. Table with the number of times each method gave the top one, top two, and top three highest EF in top 100 compounds, 1%, 3%, 5%, 10%, 15%, 20%, and 50% of the data sets. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

* Phone: +46-18-471-4394. Fax: +46-18-471-4474. E-mail: christian.skold@orgfarm.uu.se.

ACKNOWLEDGMENTS

A. K. thanks the Carl Trygger Foundation for financial support. The authors thank Dr. Luke Odell for linguistic revision of the manuscript.

REFERENCES

(1) Jorgensen, W. L. The Many Roles of Computation in Drug Discovery. *Science* **2004**, *303*, 1813−1818.

(2) Clark, D. E. What has virtual screening ever done for drug discovery? *Exp. Opin. Drug Discovery* **2008**, *3*, 841−851.

(3) Jacobsson, M.; Gäredal, M.; Schultz, J.; Karlén, A. Identification of *Plasmodium falciparum* Spermidine Synthase Active Site Binders through Structure-Based Virtual Screening. *J. Med. Chem.* **2008**, *51*, 2777−2786.

(4) Luzhkov, V. B.; Selisko, B.; Nordqvist, A.; Peyrane, F.; Decroly, E.; Alvarez, K.; Karlen, A.; Canard, B.; Åqvist, J. Virtual Screening and bioassay study of novel inhibitors for dengue virus mRNA cap (nucleoside-2O)-methyltransferase. *Bioorg. Med. Chem.* **2007**, *15*, 7795−7802.

(5) Knox, A. J. S.; Price, T.; Pawlak, M.; Golfis, G.; Flood, C. T.; Fayne, D.; Williams, D. C.; Meegan, M. J.; Lloyd, D. G. Integration of Ligand and Structure-Based Virtual Screening for the Identification of the First Dual Targeting Agent for Heat Shock Protein 90 (Hsp90) and Tubulin. *J. Med. Chem.* **2009**, *52*, 2177−2180.

(6) Lee, H. S.; Choi, J.; Kufareva, I.; Abagyan, R.; Filikov, A.; Yang, Y.; Yoon, S. Optimization of High Throughout Virtual Screening by Combining Shape-Matching and Docking Methods. *J. Chem. Inf. Model.* **2008**, *48*, 489−497.
(7) Swann, S. L.; Brown, S. P.; Muchmore, S. W.; Patel, H.; Merta, P.; Locklear, J.; Hajduk, P. J. A Unified, Probabilistic Framework for Structure- and Ligand-Based Virtual Screening. *J. Med. Chem.* **2011**, *54*, 1223−1232.
(8) Chun Meng, S.; Shen Jean, L.; Joo Chuan, T. Recent advances in computer-aided drug design. *Briefings Bioinf.* **2009**, *10*, 579−591.
(9) Johanson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons, Inc: New York, 1990.
(10) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of Protein - Ligand Interactions. Docking and scoring: Successes and Gaps. *J. Med. Chem.* **2006**, *49*, 5851−5855.
(11) Tirado-Rives, J.; Jorgensen, W. L. Contribution of Conformer Focusing to the Uncertainty in Predicting Free Eneregies for Protein-Ligand Bindning. *J. Med. Chem.* **2006**, *49*, 5880−5884.
(12) Huang, S. Y.; Grinter, S. Z.; Zou, X. Scoring funtions and their evaluation methods for protein-ligand docking: recent adavences and future directions. *Phys. Chem. Chem. Phys.* **2010**, *12*, 12899−12908.
(13) Tan, L.; Geppert, H.; Sisay, M. T.; Gütschow, M.; Bajorath, J. Integrating Structure- and Ligand-Based Virtual Screening: Comparison of Individual, Parallel, and Fused Molecular Docking and Similarity Search Calculations on Multiple Targets. *ChemMedChem* **2008**, *3*, 1566−1571.
(14) Raymond, J. W.; Jalaie, M.; Bradley, M. P. Conditional Probability: A New Fusion Method for Merging Disparate Virtual Screening Results. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 601−609.
(15) Oda, A.; Tsuchida, K.; Takakura, T.; Yamaotsu, N.; Hirono, S. Comparison of Consensus Scoring Strategies for Evaluating Computational Models of Protein-Ligand Complexes. *J. Chem. Inf. Model.* **2006**, *46*, 380−391.
(16) Baber, J. C.; Shirley, W. A.; Gao, Y.; Feher, M. The Use of Consensus Scoring in Ligand-Based Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46*, 244−288.
(17) Willet, P. Enhancing the Effectivness of Ligand-Based Virtual Screening Using Data Fusion. *QSAR Comb. Sci.* **2006**, *25*, 1143−1152.
(18) Whittle, M.; Gillet, V. J.; Willet, P. Analysis of Data Fusion Methods in Virtual Screening: Theoretical Model. *J. Chem. Inf. Model.* **2006**, *46*, 2193−2205.
(19) Whittle, M.; Gillet, V. J.; Willet, P. Analysis of Data Fusion Methods in Virtual Screening: Similarity and Group Fusion. *J. Chem. Inf. Model.* **2006**, *46*, 2206−2219.
(20) Cross, S.; Baroni, M.; Carosati, E.; Benedetti, P.; Clementi, S. FLAP: GRID Molecular Interaction Fileds in Virtual Screening. Validation using the DUD Data Set. *J. Chem. Inf. Model.* **2010**, *50*, 1442−1450.
(21) Wang, R.; Wang, S. How Does Consensus Scoring Work for Virtual Library Screening? An Idealized Comuter Experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422−1426.
(22) Krüger, D. M.; Evers, A. Comparison of Structure- and Ligand-Based Virtual Screening Protocols Considering Hit List Complementarity and Enrichment Factors. *ChemMedChem* **2010**, *5*, 148−158.
(23) Huang, N.; Scoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.
(24) Irwin, J. J. Community benchmarks for virtual screening. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 193−199.
(25) Cheeseright, T. J.; Mackey, M. D.; Melville, J. L.; Vinter, J. G. FielsScreen: Virtual Screening Using Molecular Fields. Application to the DUD Data Set. *J. Chem. Inf. Model.* **2008**, *48*, 2108−2117.
(26) Jacobsson, M.; Karlén, A. Ligand Bias of Scoring Functions in Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46*, 1334−1343.
(27) Muthas, D.; Sabnis, Y. A.; Lundborg, M.; Karlén, A. Is it possible to increse hit rates in structure-based virtual screening by pharmacophore filtering? An investigation of the advantages and pitfalls of post-filtering. *J. Mol. Graphics Modell.* **2008**, *26*, 1237−1251.
(28) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(29) 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.
(30) Kalliokoski, T.; Salo, H. S.; Lahtela-Kakkonen, M.; Poso, A. The Effect of Ligand-Based Tautomer and Protomer Prediction on Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2009**, *49*, 2742−2748.
(31) *LigPrep*, 2.4; Schrödinger, LLC: New York.
(32) Shelly, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. Epik: a software program for pKa prediction and protonation state generation for drug-like molecules. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 681−691.
(33) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissing, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(34) *Schrödinger Suite 2009 Protein Preparation Wizard*, Epik version 2.0; Schrödinger, LLC, New York, 2009. *Impact*, version 5.5, Schrödinger, LLC: New York, 2009. *Prime*, version 2.1; Schrödinger, LLC: New York, 2009.
(35) *Glide*, 5.6; Schrodinger, LLC: New York.
(36) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelly, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.
(37) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **2004**, *47*, 1750−1759.
(38) Dixon, S. L.; Smondyrev, A. M.; Knoll, E. H.; Rao, S. N.; Shaw, D. E.; Freisner, R. A. PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D databases screening: 1. Methodology and preliminary results. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 647−671.
(39) *ROCS*, 2.3.1; OpenEye Scientific Software: Santa Fe, NM.
(40) *EON*, 2.0.1; OpenEye Scientific Software: Santa Fe, NM.
(41) 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.
(42) Gund, P. Tree-Dimensional Pharmacophoric Pattern Searching. *Prog. Mol. Subcell. Biol.* **1977**, *5*, 117−143.
(43) *OMEGA*, 2.3.1; OpenEye Scientific Software: Santa Fe, NM.
(44) Boström, J.; Greenwood, J. R.; Gottfries, J. Assessing the performance of OMEGA with respect to retriving bioactive conformations. *J. Mol. Graphics Modell.* **2003**, *21*, 449−462.
(45) Diller, D. J.; Li, R. Kinases, Homology Models, and High Throughput Docking. *J. Med. Chem.* **2003**, *46*, 4638−4647.