Ligand Bias of Scoring Functions in Structure-Based Virtual Screening

Micael Jacobsson*,[†,‡] and Anders Karlén[†]

Department of Medicinal Chemistry, Faculty of Pharmacy, University of Uppsala, Box 574,
SE-751 23 Uppsala, Sweden, and Department of Chemistry, Biovitrum AB, SE-112 76 Stockholm, Sweden

Received September 16, 2005

A total of 945 known actives and roughly 10 000 decoy compounds were docked to eight different targets, and the resulting poses were scored using 10 different scoring functions. Three different score postprocessing methods were evaluated with respect to improvement of the enrichment in virtual screening. The three procedures were (i) multiple active site correction (MASC) as has been proposed by Vigers and Rizzi, (ii) a variation of MASC where corrections terms are predicted from simple molecular descriptors through PLS, PLS MASC, and (iii) size normalization. It was found that MASC did not generally improve the enrichment factors when compared to uncorrected scoring functions. For some combinations of scoring functions and targets, the enrichment was improved, for others not. However, by excluding the standard deviation from the MASC equation and transforming the scores for each target to a mean of 0 and a standard deviation of 1 (unit variance normalization), the performance was improved as compared to the original MASC method for most combinations of targets and scoring functions. Furthermore, when the molecular descriptors were fit to the mean scores over all targets and the resulting PLS models were used to predict mean scores, the enrichment as compared to the raw score was improved more often than by straightforward MASC. A high to intermediate linear correlation between the score and the number of heavy atoms was found for all scoring functions except FlexX. There seems to be a correlation between the size dependence of a scoring function and the effectiveness of PLS MASC in increasing the enrichment for that scoring function. Finally, normalization by molecular weight or heavy atom count was sometimes successful in increasing the enrichment. Dividing by the square or cubic root of the molecular weight or heavy atom count instead was often more successful. These results taken together suggest that ligand bias in scoring functions is a source of false positives in structure-based virtual screening. The number of false positives caused by ligand bias may be decreased using, for example, the PLS MASC procedure proposed in this study.

1. INTRODUCTION

Structure-based virtual screening is a well-established method for identifying binders to targets with a known structure.[1−3] The basic methodology is to dock a set of compounds into the active site of the target and use one or more scoring functions to predict the free energy of binding from the predicted binding pose. This score is used to sort the compounds, and the top scoring compounds are selected for experimental testing or further computational evaluation.

Numerous scoring functions have been published, and there are excellent reviews of docking methods and scoring functions[4−8] as well as a multitude of comparisons.[9−15] A number of methods aiming to increase the predictive power of existing scoring functions have also been described previously. These include supervised methods, using known actives to increase success in virtual screening,[16,17] as well as unsupervised methods such as consensus scoring[18,19] and mixing of ligand scores and ligand structures.[20,21]

This study was inspired by the work by Vigers and Rizzi,[22] in which the method called multiple active site correction, MASC, was introduced. Their idea was to calculate a corrected score by docking ligands to many different targets, of which only one is the intended target. From the scores for all targets, the mean score and standard deviation for each compound is calculated. The corrected score is calculated by subtracting the mean score from the raw score followed by division by the standard deviation. The mean may be considered a measure of the ligand bias, or intrinsic docking score, correlated more to ligand-specific properties such as molecular size than it is to the number and nature of target−ligand interactions.

Here, we have evaluated the use of MASC to increase enrichment in virtual screening, using a more extensive test set than in the initial study, in terms of the number of targets, compounds, and scoring functions. We have also evaluated two alternative approaches to correct for ligand bias: PLS prediction of the mean scores from molecular properties and normalization of the raw scores by different size descriptors. With this, we aim both to characterize ligand bias and to improve enrichment in unsupervised structure-based virtual screening.

* Corresponding author phone: +46 8 6972551; e-mail: micael.jacobsson@biovitrum.com.
[†] University of Uppsala.
[‡] Biovitrum AB.

10.1021/ci050407t CCC: $33.50
Published on Web 01/04/2006

Table 1. Ligand Sets Used as Known Actives

target	source	number of compounds
DHFR	Otzen	30
	Zolli-Juran	11
	PDB	4
MMP3	Ha	60
	PDB	14
AChE	Contreras	54
	PDB	9
ERα	Sippl	36
	Shi	109
	PDB	5
fXa	Matter	129
	PDB	26
COX2	Clark	305
	PDB	5
NA	Yi[a]	40
	PDB	10
CDK2	Tominaga	54
	PDB	44

[a] The stereoizer part of LigPrep was not used, since some actives are enantiomers of each other.

2. METHODS

2.1. Preparation of Ligands. A total of 10 000 random diverse compounds were used as decoys, from which the known actives were to be distinguished. The term "decoy" is used to distinguish these compounds from true inactives, known to not bind to any of the eight targets included in the study. These 10 000 compounds were selected from a database of a few hundred thousand commercially available screening compounds, using the OptiSim[23] method as implemented in ChemEnlighten,[24] with a Tanimoto radius of 0.85. The commercially available compounds were filtered according to Lipinski's rule of five[25] prior to selection. The initial 10 000 structures were converted to 3D and minimized using LigPrep.[26] LigPrep generated several different ionization states, tautomers, and stereoisomers for each compound. The final number of decoy structures was 18 851, representing 9990 compounds. The number of decoys was reduced by 10 compounds because they could not be handled by the OPLS-AA[27] force field and were, therefore, filtered away by LigPrep.

Compounds known to be active were collected from the literature[28−36] and from cocrystal structures in the Protein Data Bank (PDB). Two-dimesional structures of the known actives are available in the Supporting Information. The sources and number of compounds are shown in Table 1.

The known actives were prepared in the same way as the decoys, including the expansion of each compound to many structures through ionization states, tautomers, and stereoisomers. This resulted in a total of 10 935 compounds, represented by 21 500 structures. These structures were then docked to each target and scored.

2.2. Preparation of Targets. The protein preparation tools in the Glide[26] package were used to prepare the protein structures from PDB for docking. The Glide tools set the protonation state of the proteins, add hydrogens, and perform restrained minimization using the OPLS-AA force field. All waters were removed from the protein structures, to "simulate" virtual screening with the purpose of finding novel binders in a set of diverse random compounds. In DHFR, the cofactor was kept, and in MMP3, the catalytic Zn^{2+} ion was kept.

For each target, a number of structures from the PDB were prepared for docking, and all compounds with a known binding pose for each target were docked to all prepared structures for that target. Visual analysis of the resulting dockings was used to select the protein structure that yielded the largest number of correct dockings. This structure was used as the target for all subsequent dockings and scorings. The selected structures for each system were 1EVE[37] (AChE), 1KE7[38] (CDK2), 1PXX[39] (COX2), 1RX3[40] (DHFR), 3ERD[41] (ERα), 1IQE (fXa), 1G49[42] (MMP3), and 1L7F[43] (NA).

2.3. Docking and Scoring. All ligand structures were docked to each target, resulting in 170 200 total dockings, using Glide SP 3.5.[44,45] One single pose was kept for each structure, which was then rescored using 10 different scoring functions. These scoring functions were CScore[19,24] (FlexX score,[46] ChemScore,[47] PMF,[48] GOLD score,[49] and DOCK score[50]), GlideScore, and EModel from Glide and XScore.[51] In XScore, only the three individual scoring functions (HMScore, HPScore, and HSScore) were used, and the average score was discarded.

To generate multi-mol2 files compatible with CScore, the Maestro files output from Glide were first converted to SD files using sdconvert[26] and then from SD file format to mol2 format using Babel.[52] The multi-mol2 files and PDB files used as input to XScore were first checked with the appropriate utility supplied by the authors of XScore.

To generate the final poses, the best scoring structure for each compound according to each separate scoring function was identified. This resulted in a maximum of 10 935 poses for each target, since all compounds did not dock to all targets. For example, the binding site of ERα in its activated form is a closed, rather small hydrophobic cavity. Large ligands did not fit into this pocket. Therefore, Glide did not return any poses for that particular combination of target and ligand.

The dockings of known actives from the PDB which resulted from sorting the structures according to the Glide scoring function in the virtual screen were inspected visually. The number of compounds with a roughly correct pose for each target is as follows: AChE 2/9, CDK2 26/44, COX2 5/5, DHFR 4/4 (folate, 5-deazafolate, and 5,10-deazafolate were included as actives, even though the cocrystal structures of them are not known), ERα 5/5, fXa 19/26, MMP3 10/14, and NA 7/10. Roughly correct pose means that the overall interactions corresponded to the interactions seen in the cocrystal structures. The mean root-mean-square deviation values for all heavy atoms between the X-ray pose and the docked pose for the successfully docked known actives from the PDB were AChE, 1.6 Å; CDK2, 1.7 Å; COX2, 0.85 Å; DHFR, 1.4 Å for methotrexate; ERα, 1.0 Å; fXa, 2.0 Å; MMP3, 2.1 Å; and NA, 2.0 Å.

We defined the mean score for a particular compound as its mean score for those targets to which it was possible to dock. The standard deviation was defined in the same way. If a compound could dock to only one target, the standard deviation was set to 1.

2.4. Virtual Screening Performance Measure. The subsetting (S) was defined as the percentage of all compounds which were predicted to be actives. The prediction was always made by sorting compounds according to the score (which could be a transformed score, such as MASC score).

If a compound did not score against the target in question, it was placed last in the list. After sorting, the top *S* percent of the list were predicted as actives. A true positive (TP) was defined as a known active predicted to be active, a false positive (FP) was defined as a decoy predicted to be active, a false negative (FN) was defined as a known active not predicted to be active, and a true negative (TN) was defined as a decoy compound not predicted to be active. The enrichment factor (EF) was then defined as the proportion of true positives in the predicted actives divided by the proportion of known actives in the entire set of molecules (eq 1).

$$\mathrm{EF} = \frac{\left(\dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\right)}{\left(\dfrac{\mathrm{TP}+\mathrm{FN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}}\right)} \qquad (1)$$

Also, complete "receiver operating characteristic" (ROC) curves[53] (plotting enrichment vs subsetting) were constructed and are appended as Supporting Information.

3. RESULTS

3.1. Description of Data Sets. Using Glide, 945 known actives (see Table 1) and 9990 chemically diverse compounds were docked and scored against eight different targets. Ten different scoring functions were applied to minimize the risk of allowing the choice of scoring function to affect the conclusions. The eight targets were chosen to cover a variety of binding pockets. They were estrogen receptor α (ERα, 3ERD[41]), with a relatively small, closed, hydrophobic binding pocket; acetylcholine esterase (AChE, 1EVE[37]), with a large, flexible, open hydrophobic binding pocket with few specific interactions; cyclooxygenase-2 (COX2, 1PXX[39]), with a slightly larger and more polar binding pocket than ERα, but still closed to solvent; neuraminidase (NA, 1L7F[43]), with a partly open, rather large and polar binding pocket; dihydrofolate reductase (DHFR, 1RX3[40]), with a partly open, rather narrow binding pocket and well-defined specific interactions in the enclosed part; cyclin-dependent kinase 2 (CDK2, 1KE7[38]), with a large, open, mainly hydrophobic binding site; matrix metalloprotease 3 (MMP3, 1G49[42]), with a solvent-exposed, polar binding site and a central catalytic Zn^{2+} ion; and the serine protease factor Xa (fXa, 1IQE), also with a large, solvent-exposed and polar binding site.

The known actives were collected both from the PDB and from the literature. In some cases, compounds taken from literature studies had a high degree of structural similarity. This was particularly true for the COX2 inhibitors. One indication of the physiochemical similarity between actives and decoys is shown in Figure 1. Here, six simple molecular descriptors have been calculated for all 10 935 compounds using Sybyl;[24] molecular weight (MW), molecular volume, molar refractivity (CMR), heavy atom count (HAC), heteroatom count, and calculated ClogP. A principal component analysis (PCA) was used to project compounds in simple two-component principal space, employing SIMCA 10.0.[54] The first two principal components describe 93.9% of the variation (R^2 of 0.939) in the data set. Roughly, the first principal component describes molecular size and the second polarity. As can be seen here, the COX2 actives cluster quite tightly (Figure 1c). Compounds in the lower left quadrant, such as the ERα actives, are small and hydrophobic (Figure 1e), and compounds in the upper right quadrant, such as the fXa actives, are large and polar (Figure 1f).

Because of the structural similarity of the known actives, the actual enrichment factors are probably not indicative of the absolute enrichment values one would obtain in a true virtual screen using only random and diverse molecules. However, the decoys cover the space well, and the purpose of this study is not to judge structure-based virtual screening in absolute terms but to compare and analyze different scoring functions and score transforms.

A large set of diverse compounds is used as decoys. In the analysis, these are assumed to be true inactives. However, given normal HTS hit rates on the order of 10^{-3}, maybe as many as 10 out of the 10 000 decoys would actually show an inhibitory effect if tested versus the eight targets. This means that we may overestimate the number of false positives, but the results should still be useful for comparing different score transforms and scoring functions.

3.2. Comparison of Scoring Functions. The enrichment factors resulting from the 3% subsetting, as defined in eq 1, are shown in Table 2. For brevity, the enrichment factors are shown only for the 3% subsetting. The trends do not differ, in general, at the 1, 5, or 10% subsetting levels. See also the ROC curves appended as Supporting Information.

It should be noted that the 10 different scoring functions were treated separately. The aim of this study was not to combine information from different scoring functions, or to compare the absolute performance of different scoring functions. Nonetheless, it is informative to look across the "none" row for each target in Table 2, whereby one can compare the performance of the 10 scoring functions in terms of enrichment in virtual screening. It is apparent that the choice of scoring function is dependent on the target; hence, it is always better to have a set of known binders to validate your virtual screening approach prior to performing large-scale structure-based virtual screening. Overall, however, the best performing scoring function in this study was the Glide scoring function.[44] This could be affected by the fact that Glide has been used for docking.

As can be seen in Table 2, the enrichments for acetylcholine esterase are around 1, meaning that the virtual screening has basically failed. When inspecting the dockings of the AChE compounds from the PDB, it is apparent that most dockings to this target have failed. Hence, the poor enrichments attained for AChE are probably, in part, a failure of the docking, not the scoring part of the virtual screen. We have seen poor results in structure-based virtual screening against AChE before.[16] AChE results are not analyzed further in this study, but the dockings and scorings versus this target are still useful to form MASC correction terms for the other targets.

3.3. MASC Scoring. Enrichment factors from two different MASC procedures are shown in rows 2 and 3 for each target in Table 2. Looking across the first two rows for each target in Table 2, the most striking result is that MASC is not always beneficial. On the contrary, for example, in the case of GlideScore, MASC helped for only two of the seven targets, while for ChemScore, MASC helped for five targets. Hence, whether MASC is beneficial for enrichment in a given

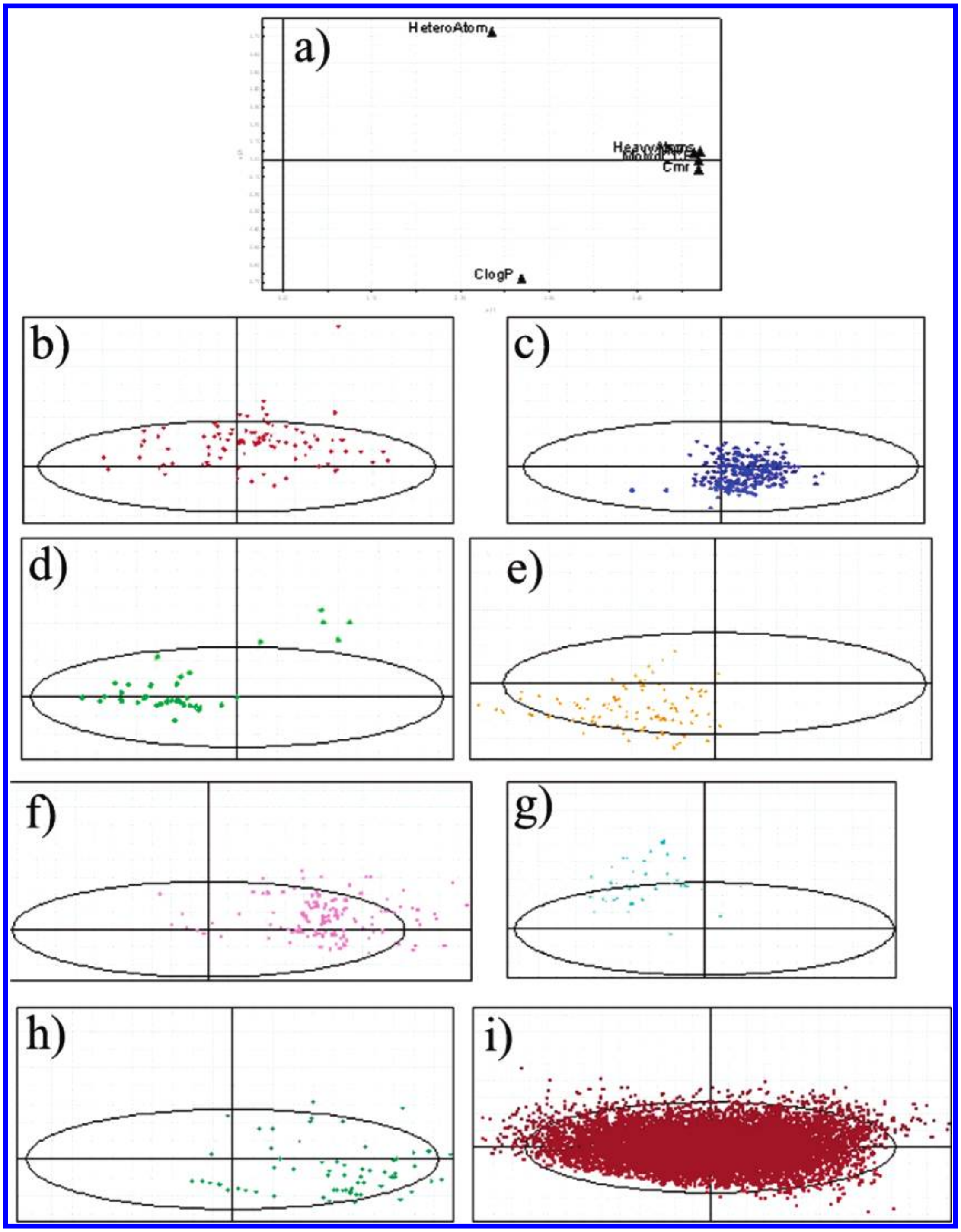

Figure 1. PCA score plots of all compounds used in the study, grouped by target. The compounds are described by six simple descriptors (HAC, MW, CMR, molecular volume, number of heteroatoms, and calculated ClogP), and the principal component space is shown in the loadings plot a. The first principal component describes size, and the second describes polarity. The score plots b−i were taken from the same PCA analysis but are shown separately for clarity. Compounds included in the score plots are (b) CDK2, (c) COX2, (d) DHFR, (e) ERα, (f) fXa, (g) NA, (h) MMP3, and (i) decoys.

case or not depends both on the scoring function and on the target.

When MASC was first proposed, the authors transformed the raw scores by subtracting the mean and then dividing by the standard deviation across the targets.[22] It is not obvious why it should be beneficial to divide by the standard deviation and thereby penalize compounds having a high variance across the targets. Accordingly, we tested only subtracting the mean from the raw score. Interestingly, we obtained a better enrichment as compared to original MASC, even though this was not true for every combination of scoring function and target. We also tried (a) to replace the average with the median, (b) to rank-transform the scores before forming averages, and (c) to not include the score versus the intended target for a specific compound in its average score over all the targets, but none of these modifications were generally beneficial (data not shown). What did help, however, was to normalize the scores from a specific target, that is, to transform the scores so that the mean was zero and the standard deviation was 1 (eq 2), before forming the mean over all the targets. As can be seen in Table 2, this modified MASC procedure (“mod. MASC”

Table 2. Enrichment Factors at 3% Subsetting for 10 Different Scoring Functions, 8 Different Targets, and 10 Score Post-Processing Procedures[a]

		scoring function									
target	transform	Glide[b]	EModel[b]	FlexX[c]	ChemScore[c]	PMF[c]	GOLD[c]	DOCK[c]	HM score[d]	HP score[d]	HS score[d]
AChE[e]	none	2.1	1.1	0	1.6	0	2.1	2.1	0.53	2.6	0
CDK2	none[f]	20	13	***21***	14	**5.1**	4.1	2.4	0.99	0.98	0.65
	MASC[g]	15	13	14	16	1.4	9.2	**8.1**	9.8	11	**13**
	mod. MASC[h]	18	**19**	19	19	4.8	**10**	7.8	12	8.5	11
	MW[i]	0.68	10	12	1.0	2.4	1.0	0	0	0	0.34
	HAC[j]	0.68	11	12	1.0	2.4	0.34	0	0	0	0.34
	SQRT(MW)[k]	14	15	19	13	4.4	2.7	1.4	0	0	0.34
	CRT(MW)[l]	20	15	19	15	4.1	3.1	1.4	0	0	0
	PLS-1[m]	21	15	**21**	14	4.8	2.7	2.0	1.0	0.34	0.34
	PLS-2[n]	21	16	21	16	4.8	2.7	2.0	1.4	1.0	0.34
	PLS-3[o]	**21**	14	21	**21**	4.4	**10**	2.4	**15**	**13**	11
	PLS-4[p]	**21**	16	21	17	4.4	3.1	2.7	1.0	1.0	1.4
COX2	none[f]	***19***	**4.7**	**9.9**	8.8	5.2	4.7	9.8	8.0	5.6	4.5
	MASC[g]	9.8	0.64	5.7	**10**	9.7	5.7	2.4	8.0	3.7	2.8
	mod. MASC[h]	14	1.6	8.7	5.5	**11**	**21**	**23**	**14**	**15**	**21**
	MW[i]	0	0.21	1.9	0	0.86	0.11	1.2	0	0	0
	HAC[j]	0	0.21	2.5	0.43	1.3	0.32	2.0	0	0	0
	SQRT(MW)[k]	3.3	0.43	6.8	3.8	2.1	1.6	5.3	0.21	0.11	0
	CRT(MW)[l]	8.8	0.86	8.3	6.1	3.2	2.6	7.8	1.9	1.2	0.21
	PLS-1[m]	13	0.64	9.7	5.2	3.3	2.4	4.9	2.7	2.3	1.1
	PLS-2[n]	13	0.54	9.8	6.3	3.4	2.7	7.0	3.8	4.0	2.0
	PLS-3[o]	15	0.75	9.8	9.8	3.7	9.0	10	3.9	4.0	4.2
	PLS-4[p]	13	0.64	9.8	6.3	3.7	2.6	7.6	3.4	3.1	1.4
DHFR	none[f]	5.8	3.0	*20*	0	*20*	1.5	3.0	0	0	0
	MASC[g]	19	**23**	19	**19**	18	3.7	3.7	**12**	5.2	**10**
	mod. MASC[h]	18	8.9	20	13	13	3.7	3.7	1.5	3.7	2.2
	MW[i]	16	15	21	8.9	26	5.2	3.0	3.7	4.4	2.2
	HAC[j]	10	10	20	5.9	27	2.2	0	1.5	2.2	1.5
	SQRT(MW)[k]	24	5.9	**21**	4.4	**28**	4.4	**4.4**	8.1	8.1	4.4
	CRT(MW)[l]	**24**	3.0	**21**	1.5	**28**	3.0	3.0	4.4	**8.9**	3.7
	PLS-1[m]	21	3.0	20	1.5	27	3.0	3.0	1.5	4.4	0
	PLS-2[n]	23	3.0	20	1.5	**28**	2.2	3.0	0.74	4.4	0
	PLS-3[o]	22	3.0	20	3.7	**28**	**6.7**	3.7	6.7	8.1	0
	PLS-4[p]	23	3.0	20	2.2	**28**	4.4	3.7	0.74	4.4	0
ERα	none[f]	*18*	12	14	14	0.22	6.2	0.67	13	10	11
	MASC[g]	20	**20**	17	14	**7.5**	3.1	0.22	7.1	4.2	7.8
	mod. MASC[h]	20	19	**18**	15	6.2	3.8	0.22	7.3	9.5	14
	MW[i]	18	14	17	21	2.2	**11**	**6.7**	9.1	9.3	6.0
	HAC[j]	14	11	18	**24**	1.6	9.5	3.1	7.5	6.9	4.6
	SQRT(MW)[k]	**21**	14	16	20	0.44	8.0	2.0	**20**	**18**	**15**
	CRT(MW)[l]	21	14	15	17	0.44	6.9	1.3	19	15	15
	PLS-1[m]	20	14	14	17	0.22	7.1	1.8	17	12	10
	PLS-2[n]	20	12	14	18	0.22	6.2	0.67	17	12	14
	PLS-3[o]	20	16	14	15	0.44	6.9	1.1	12	10	7.5
	PLS-4[p]	20	13	14	20	0.22	6.9	0.67	17	12	12
fXa	none[f]	***23***	21	**16**	**14**	**10**	**8.6**	**8.6**	**9.2**	**9.5**	**5.7**
	MASC[g]	11	9.2	9.9	4.1	0	3.2	1.5	1.5	4.3	3.2
	mod. MASC[h]	13	11	15	8.4	0	3.4	7.1	3.7	6.2	6.7
	MW[i]	1.3	7.5	4.9	0	0	0	0	0	0	0
	HAC[j]	0	6.0	2.8	0	0	0	0	0	0	0
	SQRT(MW)[k]	14	**24**	13	4.9	5.2	3.9	3.7	0	0	0
	CRT(MW)[l]	19	19	14	9.0	8.2	6.7	6.2	0.86	3.2	0.43
	PLS-1[m]	20	18	15	6.0	5.4	4.5	3.7	1.9	4.7	0.64
	PLS-2[n]	19	18	15	5.4	2.6	3.2	3.0	0.86	2.8	0.21
	PLS-3[o]	17	17	15	6.7	1.7	2.1	2.6	3.7	3.4	0.64
	PLS-4[p]	19	18	15	5.6	3.9	3.4	3.2	1.1	3.7	0.43
MMP3	none[f]	**31**	***31***	**28**	**28**	**3.6**	**9.4**	**12**	**19**	**22**	**14**
	MASC[g]	0.90	0.45	0.45	1.4	0	1.4	1.8	0	0.90	0
	mod. MASC[h]	14	24	7.2	7.2	0.45	1.8	0	1.8	6.7	8.5
	MW[i]	0	5.8	1.8	2.3	0	0	0	0	0	0
	HAC[j]	0	4.9	1.8	1.8	0	0	0	0	0	0
	SQRT(MW)[k]	9.0	24	15	15	0	2.7	3.1	0	0	0
	CRT(MW)[l]	23	31	14	20	0.90	4.9	6.7	11	11	1.4
	PLS-1[m]	27	31	27	20	0.90	3.1	3.1	13	17	5.4
	PLS-2[n]	27	30	27	20	0	2.7	2.3	10	18	4.0
	PLS-3[o]	28	31	27	15	2.3	2.3	1.8	3.6	11	3.1
	PLS-4[p]	27	30	27	19	0	2.3	1.8	11	16	4.9

Table 2 (Continued)

		scoring function									
target	transform	Glide[a]	EModel[a]	FlexX[b]	ChemScore[b]	PMF[b]	GOLD[b]	DOCK[b]	HM score[c]	HP score[c]	HS score[c]
NA	none[f]	18	*27*	24	6.6	21	7.3	11	0	0	0
	MASC[g]	5.3	27	17	23	27	5.3	14	**23**	**25**	**24**
	mod. MASC[h]	21	**31**	**31**	**26**	**32**	11	**19**	16	17	19
	MW[i]	13	28	25	9.3	25	9.3	12	0	0	0
	HAC[j]	10	28	25	7.3	25	8.6	10	0	0	0
	SQRT(MW)[k]	**25**	30	27	14	27	11	17	1.3	0	0
	CRT(MW)[l]	24	29	27	13	26	11	17	0.67	0.67	0.67
	PLS-1[m]	23	29	24	13	26	11	17	0	1.3	0.67
	PLS-2[n]	23	29	24	12	26	11	17	0	2.0	0.67
	PLS-3[o]	22	27	24	21	26	10	16	13	18	15
	PLS-4[p]	23	29	24	13	26	**12**	17	0	1.3	0.67

[a] The best scoring function for each target is italicized. The best score transform for each combination of target and scoring function is shown in bold. [b] Glide. [c] CScore. [d] XScore. [e] Since the result for AChE is basically random, it is not included in the analysis. [f] Raw, untransformed score. [g] MASC as defined by Vigers and Rizzi. [h] MASC as defined by eqs 2 and 3. [i] Size normalization using MW. [j] Size normalization using HAC.[k] Size normalization using square root of MW. [l] Size normalization using cubic root of MW. [m] PLS MASC using a PLS model being a least-squares fit between molecular weight and mean score. [n] PLS MASC using a PLS model being a least-squares fit between HAC and mean score. [o] PLS MASC using a four-component PLS model with six different molecular descriptors—MW, HAC, molecular volume, CMR, ClogP, and number of heteroatoms. [p] PLS MASC using a one-component PLS model with three different molecular descriptors—MW, HAC, and molecular volume.

in the table) most often performed better than the original MASC procedure. Modified MASC is summarized in eqs 2 and 3 below (μ, σ, and s are the mean, standard deviation, and score, respectively).

$$s_{\mathrm{UN}}^{i,j} = \frac{s_{\mathrm{raw}}^{i,j} - \mu_{\mathrm{raw}}^{j}}{\sigma_{\mathrm{raw}}^{j}} \qquad (2)$$

$$s_{\mathrm{MASC}}^{i,j} = s_{\mathrm{UN}}^{i,j} - \mu_{\mathrm{UN}}^{i} \qquad (3)$$

Here, i denotes compound and j denotes target; hence, μ_{UN}^{i} means the average of all scores for compound i versus all eight targets, where the scores have been normalized before being averaged, while μ_{raw}^{j} means the average of the raw scores for all compounds that dock to target j.

3.4. Size Normalization. The basic idea behind MASC as described by Vigers and Rizzi is that the mean score for a compound scored against multiple targets should mirror the "nonspecific" part of that compound's score against its intended target. Scoring functions are generally correlated with molecular size,[55] and hence, large compounds often show up as false positives in structure-based virtual screens. This correlation probably comes from the fact that most scoring functions include attractive terms summing over all atoms in the compounds being scored. To counter this size-dependent contribution to the score, effects counteracting binding, such as mismatched polar interactions, buried polar atoms, or hydrophobic atoms exposed to the solvent upon binding, need to be properly modeled. These effects are, however, difficult to include properly in fast force-field-based or empirical-scoring functions, relying on continuous solvent models and a single binding conformation.

Pan et al. suggested[55] that the bias toward scoring larger molecules higher could partly be compensated for by normalizing the scores by size. They used large diverse sets of molecules docked and scored to HIV-1 integrase (no known binders included) to show that the molecular weight distribution of the predicted binders are more similar to the starting distribution of the entire set of molecules if one normalizes the scores by the square or cubic root of the number of heavy atoms. We, therefore, normalized the scores in this study using the MW, HAC, CMR, molecular volume, and the square and cubic roots of MW and HAC. Parts of the results are shown in rows 4−7 for each target in Table 2.

As can be seen in Table 2, simple normalization by size sometimes performed as good as or better than modified MASC. It is also evident that, as hypothesized by Pan et al., dividing by the square or cubic root of MW or HAC (data not shown for HAC) was often better than using MW or HAC. For fXa and MMP3, the results were not improved by size normalization, probably because these binding sites and the known actives were larger (see Figure 1). MASC did not improve the enrichments for these targets either (see Table 2). The enrichments for DHFR, ERα, and NA were all improved by size normalization, in accordance with the smaller binding pockets and smaller known actives. The results for these targets were also improved by modified MASC. Hence, for these targets and known actives, MASC probably acts through size normalization. In the case of COX2, on the other hand, the enrichment was not improved by size normalization, but by MASC. So, here, MASC might correct primarily for some other type of "ligand-dependent overscoring".

3.5. PLS MASC. If MASC is able to capture the ligand-dependent part of the score of a compound, it might be possible to predict this ligand bias of a scoring function from molecular descriptors. Therefore, we constructed a number of PLS[56] models, correlating the mean scores for each compound when scored against all eight targets to simple molecular descriptors. This was done for all 10 scoring functions separately. These "predicted MASC terms" were subtracted from the actual score to form an alternative MASC score, PLS MASC (see Figure 2). More complex PLS models, based on 192 different molecular descriptors, either directly or by using hierarchical PLS, were also tested, but the results were not improved compared to the simpler models. The enrichment factors resulting from this procedure, using four different PLS models, are shown in the four last rows for each target in Table 2.

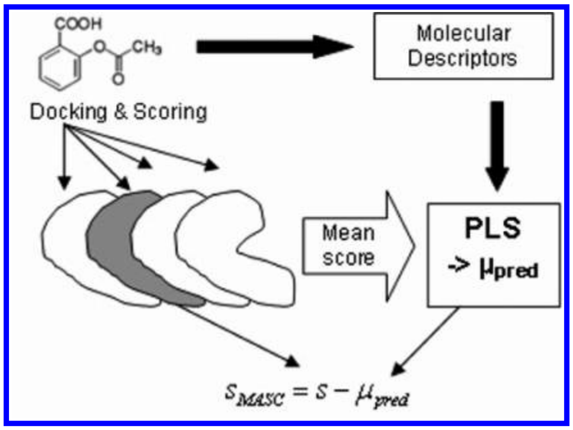

Figure 2. Schematic presentation of how the PLS MASC scores are calculated. The intended target is shown in gray.

From Table 2, it is evident that PLS MASC often produced a better enrichment than original MASC. This might be due to compounds that obtained a higher or lower mean score than what was warranted by their structures alone. The mean scores of these outlier compounds mirror actual attractive or repulsive interactions with some of the targets used to form the mean. Predicting the mean using PLS models instead might more closely reflect the ligand bias part of the scores of such compounds.

4. DISCUSSION

This study partly highlights the problem of ligand bias in scoring functions and partly tries to approach the problem. The test we have chosen is a rather simplistic approach to virtual screening. In a real case, if one knows more about the target for which novel binders are sought, this knowledge can be used to filter what compounds to dock and score and to create pharmacophoric features to help in removing obvious false positives (e.g., Brenk et al.[57]). However, structure-based screening of novel targets, where basically nothing is known of what to expect from true hits, is something we and certainly others do and would like to do as successfully as possible. In other words, today, if one would want to find novel binders of, for example, CDK-2, one could probably do much better using all the structural and medicinal chemistry information available about this target than simply docking and scoring hundreds of thousands of random and diverse compounds. The ligand bias of scoring functions would then most likely not be a significant problem. However, if the target is novel, from, for example, structural genomics, docking and scoring is more important for finding true binders, and ligand bias is accordingly a more serious problem.

It seems as if the most important molecular property for ligand bias is molecular size. All 10 scoring functions in this study are more or less correlated to molecular size. The r^2 values, when fitting the scores from each target and scoring function to one simple descriptor, HAC, are shown in Table 3. All scoring functions except for FlexX show intermediate to high correlation. This is particularly true for DOCK score and the three XScore scoring functions. Also, when constructing different PLS models correlating the score to molecular descriptors during the development and evaluation of PLS MASC, the MW, HAC, and molecular volume were most significant, even when including many molecular descriptors.

The DOCK scoring function in CScore is implemented as a simple sum over all the ligand atoms of vdW and Coulomb interactions (eq 4[58]), and the high correlation with the number of heavy atoms evident in Table 3 is easy to understand.

$$D_{\text{score}} = \sum_{\text{lig}} \sum_{\text{prot}} \left(\frac{A}{r_{ij}^{12}} - \frac{B}{r_{ij}^{6}} + 332.0 \frac{q_i q_j}{D r_{ij}} \right) \qquad (4)$$

In XScore, the authors have added a "deformation term" (eq 5[51]) that is supposed to account for the loss of entropy upon binding, but this is obviously not enough to counter the nonspecific score contribution from simply having more atoms.

$$\Delta G_{\text{bind}} = \Delta G_{\text{vdw}} + \Delta G_{\text{H-bond}} + \Delta G_{\text{deformation}} + \Delta G_{\text{hydrophobic}} + \Delta G_0 \quad (5)$$

The FlexX scoring function (eq 6[59]) contains a sum over all atoms ("lipophilic contacts") but also a penalty term proportional to the number of rotatable bonds.

$$\Delta G = \Delta G_0 + \Delta G_0 N_{\text{rot}} + \Delta G_{\text{hb}} \sum_{\text{neutral H-bonds}} f(\Delta R, \Delta\alpha) + \Delta G_{\text{io}} \sum_{\text{ionic int.}} f(\Delta R, \Delta\alpha) + \Delta G_{\text{aro}} \sum_{\text{aro int.}} f(\Delta R, \Delta\alpha) + \Delta G_{\text{lipo}} \sum_{\text{lipo.cont.}} f^{x}(\Delta R) \quad (6)$$

Maybe this term, or the design of the lipophilic contacts term, explains the lack of correlation with HAC for the FlexX score. Interestingly, the DOCK and XScore scoring functions more often benefit from "correction" with PLS-predicted MASC terms predicted from molecular size descriptors than the FlexX scoring function (FlexX, DOCK, HMScore, HPScore, and HSScore columns of Table 2).

Note that, for each target, the number of decoys is between 2 and 3 magnitudes higher than the number of actives. In a series of close analogues, where all compounds bind to a target, the experimental affinity is heavily correlated to the size of the molecule, as long as there are no steric clashes. However, for a set of diverse random molecules, this is not true. The observed correlation between size and score is accordingly a source of false positives in structure-based virtual screening. To decrease this correlation, a scoring function probably needs to correctly incorporate more subtle physical effects affecting the experimental binding energy. These could include polar groups in the ligand or the protein being desolvated upon binding but not finding a matching polar interaction in the complex; hydrophobic patches of the ligand exposed to the solvent upon binding; entropic effects such as hydrophobic interactions and a loss of rotational and translational entropy; and so forth. Also, scoring functions often use soft vdW potentials to avoid a too-high repulsion from clashes. The need for soft potentials is caused by the necessity to model the target as a rigid body, while trying to be able to adequately score as general true binders as possible. However, true clashes are detrimental for binding and would result in a positive binding energy.

Table 3. Correlation Coefficient (r^2) between Heavy Atom Count (HAC) and the Scores vs the Eight Targets for All 10 Scoring Functions

	mean	AChE	CDK2	COX2	DHFR	ERα	fXa	MMP3	NA
GlideScore	0.43	0.60	0.30	0.11	0.24	0.024	0.43	0.19	0.18
EModel	0.31	0.33	0.44	0.096	0.17	0.089	0.36	0.20	0.18
FlexX	0.0040	0.0016	0.0026	0.0013	0.0022	0.00083	0.0039	0.0012	0.0019
ChemScore	0.49	0.58	0.33	0.55	0.35	0.42	0.39	0.40	0.19
PMF	0.34	0.52	0.096	0.25	0.16	0.061	0.050	0.22	0.22
GOLD	0.37	0.39	0.32	0.57	0.18	0.41	0.30	0.23	0.22
DOCK	0.73	0.70	0.70	0.89	0.33	0.87	0.43	0.44	0.58
HMScore	0.61	0.70	0.45	0.64	0.50	0.54	0.56	0.63	0.50
HPScore	0.68	0.76	0.53	0.73	0.41	0.64	0.63	0.57	0.56
HSScore	0.59	0.71	0.51	0.65	0.38	0.52	0.59	0.50	0.46

In other words, maximizing the enrichment in virtual screening is not the same thing as maximizing the correlation with known experimental affinities. Correlation between score and molecular size is not a problem, but a necessity, if only known actives are docked and scored. It is a problem, however, if the primary aim is to distinguish binders from nonbinders, and the correlation remains even if you dock and score a set of compounds where maybe at most 1% are true binders. Scoring functions intended for virtual screening should be able to separate binders from nonbinders. Ranking the binders could very well be handled by a separate scoring function. If the enrichment of a virtual screening procedure is high enough, the predicted binders can even be verified and ranked experimentally. Postprocessing of the scores in the ways evaluated in this study might be one way to tailor existing scoring functions better to the task of finding true binders in a large set of mostly inactive compounds.

5. CONCLUSIONS

To summarize, we have docked and scored 945 known actives and 9990 decoy compounds to eight different targets, with one docking program but using 10 different scoring functions. We have tested three different unsupervised postprocessing procedures of scores to try to improve the enrichment in structure-based virtual screening: (i) different variations of MASC, (ii) PLS models capturing the ligand bias part of the scores, PLS MASC, and (iii) size normalization. These procedures have in common that they try to correct for the bias toward large molecules observed for many existing scoring functions (Pan et al.,[55] Table 3, and personal experience).

One conclusion from this study is that MASC as introduced by Vigers and Rizzi does not improve the performance of structure-based virtual screening, in general, for the data set used here. For some combinations of targets and scoring functions, it does; for others, it does not. Modifying the original equations to not include the standard deviation over the targets and performing normalization of the scores for each target prior to MASC (eq 2 and 3) increases the performance. The enrichment is not further improved by replacing mean by median, by rank-transforming the scores prior to forming averages, or, to any significant degree, by excluding the intended target from the average over all the targets. Even for the best performing MASC procedure identified in this study, the enrichment is still not improved for all the targets (Table 2).

For DHFR, ERα, and NA, size normalization using MW or HAC increased the enrichment for most scoring functions, as compared to the raw score. By instead using the square or cubic root of the molecular weight, the enrichment was improved also for CDK2 for some scoring functions. The results for MMP3 and fXa, however, were drastically deteriorated, less so when using the square or cubic roots. This is probably because the known actives for these two targets were larger than for the other targets (see Figure 1), which in turn was a consequence of these proteins having large, solvent-exposed active sites. The known actives used for MMP3 and fXa were, however, not outliers compared to the decoy compounds docked and scored in this study. Evidently, the bias toward larger molecules shown by the scoring functions studied here (Table 3) was less detrimental for identifying large true binders. When introducing a penalty for being large through size normalization, this needs to be considered.

When PLS MASC was performed as is illustrated in Figure 2, the resulting improvement in enrichment was often more pronounced than for the best MASC procedure in Table 2. Perhaps as important, the reduction in enrichment for MMP3 and fXa was not as drastic. This procedure probably works better as compared to straightforward MASC since it captures the ligand bias of the scoring function in question in a more robust way. However, this method also failed to improve the enrichment for certain combinations of targets and scoring functions.

The PLS models in Table 2 are very simple; two are only least-squares fits of docking scores to MW (PLS-1) or HAC (PLS-2), while the PLS-3 model includes some polarity information (calculated ClogP and the number of heteroatoms), and the PLS-4 model is basically a least-squares fit to a size descriptor which is a weighted average of three other heavily correlated size descriptors. The correlations observed in Table 3, together with the fact that the enrichments were improved by correcting raw scores using such PLS models, show that molecular size is an important factor for a compound to obtain a high score and that this often deteriorates the results of a structure-based virtual screen.

ACKNOWLEDGMENT

We thank Ludvig Holmberg (Uppsala University) for assistance in carrying out the docking and scoring calculations, Mats Kihlén (Biovitrum AB) and Shane Peterson (Uppsala University) for advice and proof reading, and Dr. Robert D. Clark (Tripos Inc.) for providing the non-PDB ligand structures for COX2. Financial support was received from the Swedish Foundation for Strategic Research (SSF).

Supporting Information Available: Two-dimensional representations of all actives used in this study. ROC curves for subsettings from 0.05% to 10.0%, representing all enrichments in this paper. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Rastelli, G.; Pacchioni, S.; Sirawaraporn, W.; Sirawaraporn, R.; Parenti, M. D.; Ferrari, A. M. Docking and database screening reveal new classes of Plasmodium falciparum dihydrofolate reductase inhibitors. *J. Med. Chem.* **2003**, *46*, 2834−2845.

(2) Doman, T. N.; McGovern, S. L.; Witherbee, B. J.; Kasten, T. P.; Kurumbail, R.; Stallings, W. C.; Connolly, D. T.; Shoichet, B. K. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J. Med. Chem.* **2002**, *45*, 2213−2221.

(3) Mozziconacci, J. C.; Arnoult, E.; Bernard, P.; Do, Q. T.; Marot, C.; Morin-Allory, L. Optimization and validation of a docking-scoring protocol; application to virtual screening for COX-2 inhibitors. *J. Med. Chem.* **2005**, *48*, 1055−1068.

(4) 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.

(5) Barril, X.; Hubbard, R. E.; Morley, S. D. Virtual screening in structure-based drug discovery. *Mini Rev. Med. Chem.* **2004**, *4*, 779−791.

(6) Alvarez, J. C. High-throughput docking as a source of novel drug leads. *Curr. Opin. Chem. Biol.* **2004**, *8*, 365−370.

(7) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862−865.

(8) Good, A. Structure-based virtual screening protocols. *Curr. Opin. Drug Discovery Dev.* **2001**, *4*, 301−307.

(9) Perola, E.; Walters, W. P.; Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins* **2004**, *56*, 235−249.

(10) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035−1042.

(11) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.

(12) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L., III. Assessing scoring functions for protein−ligand interactions. *J. Med. Chem.* **2004**, *47*, 3032−3047.

(13) Schulz-Gasch, T.; Stahl, M. Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. *J. Mol. Model.* **2003**, *9*, 47−57.

(14) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of docking performance: comparative data on docking algorithms. *J. Med. Chem.* **2004**, *47*, 558−565.

(15) Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46*, 2287−2303.

(16) Jacobsson, M.; Liden, P.; Stjernschantz, E.; Bostrom, H.; Norinder, U. Improving structure-based virtual screening by multivariate analysis of scoring data. *J. Med. Chem.* **2003**, *46*, 5781−5789.

(17) Terp, G. E.; Johansen, B. N.; Christensen, I. T.; Jorgensen, F. S. A new concept for multidimensional selection of ligand conformations (MultiSelect) and multidimensional scoring (MultiScore) of protein−ligand binding affinities. *J. Med. Chem.* **2001**, *44*, 2333−2343.

(18) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.

(19) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus scoring for ligand/protein interactions. *J. Mol. Graphics Modell.* **2002**, *20*, 281−295.

(20) Klon, A. E.; Glick, M.; Davies, J. W. Combination of a naive Bayes classifier with consensus scoring improves enrichment of high-throughput docking results. *J. Med. Chem.* **2004**, *47*, 4356−4359.

(21) Klon, A. E.; Glick, M.; Thoma, M.; Acklin, P.; Davies, J. W. Finding more needles in the haystack: A simple and efficient method for improving high-throughput docking results. *J. Med. Chem.* **2004**, *47*, 2743−2749.

(22) Vigers, G. P.; Rizzi, J. P. Multiple active site corrections for docking and virtual screening. *J. Med. Chem.* **2004**, *47*, 80−89.

(23) Clark, R. D. OptiSim: An Extended Dissimilarity Selection Method for Finding Diverse Representative Subsets. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1181−1188.

(24) Tripos Inc., 1699 South Hanley Road, St. Louis, MO 63144-2319.

(25) 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.

(26) Schrödinger LLC., 120 West 45th Street, 32nd Floor, New York, NY 10036-4041.

(27) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.

(28) Otzen, T.; Wempe, E. G.; Kunz, B.; Bartels, R.; Lehwark-Yvetot, G.; Hansel, W.; Schaper, K. J.; Seydel, J. K. Folate-synthesizing enzyme system as target for development of inhibitors and inhibitor combinations against *Candida albicans*−synthesis and biological activity of new 2,4-diaminopyrimidines and 4′-substituted 4-aminodiphenyl sulfones. *J. Med. Chem.* **2004**, *47*, 240−253.

(29) Zolli-Juran, M.; Cechetto, J. D.; Hartlen, R.; Daigle, D. M.; Brown, E. D. High throughput screening identifies novel inhibitors of *Escherichia coli* dihydrofolate reductase that are competitive with dihydrofolate. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 2493−2496.

(30) Ha, S.; Andreani, R.; Robbins, A.; Muegge, I. Evaluation of docking/scoring approaches: a comparative study based on MMP3 inhibitors. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 435−448.

(31) Contreras, J. M.; Parrot, I.; Sippl, W.; Rival, Y. M.; Wermuth, C. G. Design, synthesis, and structure−activity relationships of a series of 3-[2-(1-benzylpiperidin-4-yl)ethylamino]pyridazine derivatives as acetylcholinesterase inhibitors. *J. Med. Chem.* **2001**, *44*, 2707−2718.

(32) Sippl, W. Receptor-based 3D QSAR analysis of estrogen receptor ligands−merging the accuracy of receptor-based alignments with the computational efficiency of ligand-based methods. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 559−572.

(33) Matter, H.; Defossa, E.; Heinelt, U.; Blohm, P. M.; Schneider, D.; Muller, A.; Herok, S.; Schreuder, H.; Liesum, A.; Brachvogel, V.; Lonze, P.; Walser, A.; Al-Obeidi, F.; Wildgoose, P. Design and quantitative structure−activity relationship of 3-amidinobenzyl-1H-indole-2-carboxamides as potent, nonchiral, and selective inhibitors of blood coagulation factor Xa. *J. Med. Chem.* **2002**, *45*, 2749−2769.

(34) Shi, L. M.; Fang, H.; Tong, W.; Wu, J.; Perkins, R.; Blair, R. M.; Branham, W. S.; Dial, S. L.; Moland, C. L.; Sheehan, D. M. QSAR models using a large diverse set of estrogens. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 186−195.

(35) Clark, R. D. Boosted leave-many-out cross-validation: the effect of training and test set diversity on PLS statistics. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 265−275.

(36) Tominaga, Y.; Jorgensen, W. L. General model for estimation of the inhibition of protein kinases using Monte Carlo simulations. *J. Med. Chem.* **2004**, *47*, 2534−2549.

(37) Kryger, G.; Silman, I.; Sussman, J. L. Structure of acetylcholinesterase complexed with E2020 (Aricept): implications for the design of new anti-Alzheimer drugs. *Struct. Fold. Des.* **1999**, *7*, 297−307.

(38) Bramson, H. N.; Corona, J.; Davis, S. T.; Dickerson, S. H.; Edelstein, M.; Frye, S. V.; Gampe, R. T., Jr.; Harris, P. A.; Hassell, A.; Holmes, W. D.; Hunter, R. N.; Lackey, K. E.; Lovejoy, B.; Luzzio, M. J.; Montana, V.; Rocque, W. J.; Rusnak, D.; Shewchuk, L.; Veal, J. M.; Walker, D. H.; Kuyper, L. F. Oxindole-based inhibitors of cyclin-dependent kinase 2 (CDK2): design, synthesis, enzymatic activities, and X-ray crystallographic analysis. *J. Med. Chem.* **2001**, *44*, 4339−4358.

(39) Rowlinson, S. W.; Kiefer, J. R.; Prusakiewicz, J. J.; Pawlitz, J. L.; Kozak, K. R.; Kalgutkar, A. S.; Stallings, W. C.; Kurumbail, R. G.; Marnett, L. J. A novel mechanism of cyclooxygenase-2 inhibition involving interactions with Ser-530 and Tyr-385. *J. Biol. Chem.* **2003**, *278*, 45763−45769.

(40) Sawaya, M. R.; Kraut, J. Loop and subdomain movements in the mechanism of *Escherichia coli* dihydrofolate reductase: crystallographic evidence. *Biochemistry* **1997**, *36*, 586−603.

(41) Shiau, A. K.; Barstad, D.; Loria, P. M.; Cheng, L.; Kushner, P. J.; Agard, D. A.; Greene, G. L. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell* **1998**, *95*, 927−937.

(42) Natchus, M. G.; Bookland, R. G.; De, B.; Almstead, N. G.; Pikul, S.; Janusz, M. J.; Heitmeyer, S. A.; Hookfin, E. B.; Hsieh, L. C.; Dowty, M. E.; Dietsch, C. R.; Patel, V. S.; Garver, S. M.; Gu, F.; Pokross, M. E.; Mieling, G. E.; Baker, T. R.; Foltz, D. J.; Peng, S. X.; Bornes, D. M.; Strojnowski, M. J.; Taiwo, Y. O. Development of new hydroxamate matrix metalloproteinase inhibitors derived from functionalized 4-aminoprolines. *J. Med. Chem.* **2000**, *43*, 4948−4963.

(43) Smith, B. J.; McKimm-Breshkin, J. L.; McDonald, M.; Fernley, R. T.; Varghese, J. N.; Colman, P. M. Structural studies of the resistance of influenza virus neuramindase to inhibitors. *J. Med. Chem.* **2002**, *45*, 2207−2212.

(44) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, 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.

(45) 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.

(46) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.

(47) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.

(48) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein−ligand interactions: a simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791−804.

(49) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.

(50) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule−ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.

(51) Wang, R. X.; Lai, L. H.; Wang, S. M. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 11−26.

(52) Walters, P.; Stahl, M. *BABEL*. http://www.ccl.net/cca/software/UNIX/babel/index.shtml.

(53) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O. Virtual screening workflow development guided by the “receiver operating characteristic” curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* **2005**, *48*, 2534−2547.

(54) Umetrics AB, Tvistevägen 48, Box 7960, SE-907 19 Umeå, Sweden.

(55) Pan, Y.; Huang, N.; Cho, S.; MacKerell, A. D., Jr. Consideration of molecular weight during compound selection in virtual target-based database screening. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 267−272.

(56) Wold, S.; Johansson, E.; Cocchi, M. PLS−Partial least-squares projections to latent structures. In *3D QSAR in Drug Design*; ESCOM: Leiden, The Netherlands, 1993; pp 523−550.

(57) Brenk, R.; Naerum, L.; Gradler, U.; Gerber, H. D.; Garcia, G. A.; Reuter, K.; Stubbs, M. T.; Klebe, G. Virtual screening for submicromolar leads of tRNA-guanine transglycosylase based on a new unexpected binding mode detected by crystal structure analysis. *J. Med. Chem.* **2003**, *46*, 1133−1143.

(58) *CScore Manual, Tripos Bookshelf*, version 7.1; Tripos Inc.: St. Louis, MO.

(59) *FlexX Manual, Tripos Bookshelf*, version 7.1; Tripose Inc.; St. Louis, MO.

CI050407T