JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Potential and Limitations of Ensemble Docking

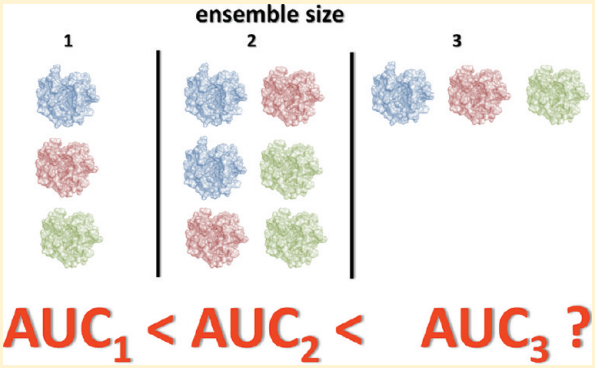
Oliver Korb,*[,†] Tjelvar S. G. Olsson,[†] Simon J. Bowden,[†] Richard J. Hall,[‡] Marcel L. Verdonk,[‡] John W. Liebeschuetz,[†] and Jason C. Cole[†]

[†]Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, U.K.

[‡]Astex Pharmaceuticals, 436 Cambridge Science Park, Milton Road, Cambridge CB4 0QA, U.K.

Supporting Information

ABSTRACT: A major problem in structure-based virtual screening applications is the appropriate selection of a single or even multiple protein structures to be used in the virtual screening process. A priori it is unknown which protein structure(s) will perform best in a virtual screening experiment. We investigated the performance of ensemble docking, as a function of ensemble size, for eight targets of pharmaceutical interest. Starting from single protein structure docking results, for each ensemble size up to 500 000 combinations of protein structures were generated, and, for each ensemble, pose prediction and virtual screening results were derived. Comparison of single to multiple protein structure results suggests improvements when looking at the performance of the worst and the average over all single protein structures to the performance of the worst and average over all protein ensembles of size two or greater, respectively. We identified several key factors affecting ensemble docking performance, including the sampling accuracy of the docking algorithm, the choice of the scoring function, and the similarity of database ligands to the cocrystallized ligands of ligand-bound protein structures in an ensemble. Due to these factors, the prospective selection of optimum ensembles is a challenging task, shown by a reassessment of published ensemble selection protocols.



INTRODUCTION

In the last decades the importance of being able to model protein flexibility has been widely recognized.[1−5] Protein−ligand docking applications, in particular, require efficient ways of modeling induced fit effects.[6] Depending on the amount of protein flexibility,[7,8] the use of soft potentials[9] or the rearrangement of a few side chains in the binding site may be sufficient.[10−17] However, for some targets major backbone movements are observed, in which case full receptor flexibility in the docking calculation may be needed.[18−21] The effect of protein flexibility on pose prediction and virtual screening performance has been studied extensively in the past.[22−28] Some of these approaches treat flexibility explicitly, allowing extra degrees of freedom in the search space to perform direct changes of the binding site conformation. In contrast to explicitly modeling flexibility, so-called *ensemble docking* methodologies make use of a limited number of discrete protein conformations. According to Rueda et al.,[29] it is usually sufficient to represent protein flexibility nonredundantly with less than 100 protein conformations.

Ensemble docking can be simulated by docking a ligand sequentially into all ensemble protein structures and post-processing the single protein structure results. While the docking time for this approach scales linearly with the number of protein structures constituting the ensemble, specific algorithms have been designed to search these ensembles time-efficiently.[30−34] Several studies have been published on constructing ensembles out of multiple experimentally determined X-ray[9,35−41] and NMR[42,43] protein structures or a combination of both.[44−46] Other sources for ensembles include computationally derived protein conformations from molecular dynamics simulations,[45,47−51] normal-mode analysis,[52] or homology models.[53] Most of these studies conclude that the use of multiple protein structures is beneficial in pose prediction and virtual screening experiments.

While the benefit of using multiple protein structures in docking has already been demonstrated in many of the studies mentioned above, we were interested in certain aspects of ensemble docking. First of all, we were interested in the performance of all possible ensembles that can be constructed out of a given set of protein structures. This analysis was carried out for eight targets of pharmaceutical relevance, with three different scoring functions available in GOLD.[54,55] While this data formed the basis for the identification of target and scoring function dependencies, it was additionally used to assess the performance of ensemble selection protocols. Unlike other studies, we have investigated the chemotype enrichment problem[56] and also investigated the influence of sampling accuracy on the ensemble docking performance.

Received: December 12, 2011
Published: April 8, 2012

   dx.doi.org/10.1021/ci2005934 | J. Chem. Inf. Model. 2012, 52, 1262−1274

MATERIALS AND METHODS

Targets. For the assessment of pose prediction and virtual screening performance, we used eight targets which were available in both the Astex non-native set[22] and the *Directory of Useful Decoys* (DUD):[57] acetylcholine esterase (*ache*), aldose reductase (*alr2*), cyclin dependent kinase 2 (*cdk2*), dihydrofolate reductase (*dhfr*), factor Xa (*fxa*), neuraminidase (*na*), p38 MAP kinase (*p38*), and phosphodiesterase 5A (*pde5*).

Docking. All experiments were carried out using GOLD version 5.0. Results were obtained for the three scoring functions ASP,[58] ChemScore,[54,59,60] and ChemPLP,[61] the latter being what we recommend for virtual screening experiments. GOLD maximizes the score, thus higher scoring function values are better.

Pose Prediction. For each target, between 1 and 20 diverse ligands were extracted from the Astex non-native set.[22] The respective protein data bank (PDB)[62] codes can be found in Table 1. Protein structures in this test set are superimposed to a common reference structure given in the Astex diverse set.[63] Consequently, for each target, a common binding site definition based on the ligand as given in the reference protein structure was used. All protein residues within 6 Å of any heavy atom of this ligand were considered as part of the binding site. Before each docking experiment, an unbiased ligand input geometry was generated using Corina.[64] For each Corina-generated ligand, 25 docking runs were performed using the following settings. The number of genetic algorithm runs was set to 10, the value for autoscale was 1.0, early termination was turned off, and the relative ligand energy option was activated. For each of the 25 runs the heavy-atom rmsd (*root-mean-square deviation* of the superimposed atomic coordinates) of the top-ranked solution was calculated with respect to the experimentally observed ligand conformation as given in the Astex non-native set. Predictions with an rmsd lower than 2 Å were classified as successful. In all docking experiments, the native protein conformation was excluded and the ligand was docked into all remaining protein structures from the Astex non-native set of the target.

Table 1. Pose Prediction Test Set

target	Astex diverse set PDB code	PDB codes of ligands used for pose prediction
acetylcholine esterase	1gpk	1dx6, 1e66, 1vot
aldose reductase	1t40	1iei, 1pwl, 1t40, 1z89, 2fzd, 2ikg, 2ikh, 2iki, 2is7
cyclin-dependent kinase 2	1ke5	1aq1, 1ckp, 1di8, 1e1x, 1e9h, 1fvt, 1jsv, 1oiq, 1p2a, 1pxj, 1vyz, 1w0x, 1y8y, 1y91, 1ykr, 2b54, 2btr, 2c68, 2c6i, 2duv
dihydrofolate reductase	1s3v	1dhf, 1hfr, 1mvs, 1s3v
factor Xa	1lpz	1ezq, 1lpz, 1mq6, 1xka, 2bok, 2fzz
neuraminidase	1l7f	1f8b, 1iny, 1l7f, 1xoe, 2qwk
p38 MAP kinase	1ywr	1a9u, 1bl7, 1kv1, 1m7q, 1ouy, 1oz1, 1w83, 1zzl, 2bak, 2gtn
phosphodiesterase 5A	1xoz	1xoz

Virtual Screening. The virtual screening experiments used the same settings as the pose prediction experiments except that only a single docking experiment with 20 GA runs per ligand was performed. The active and decoy sets were taken from the DUD data set. The geometry of each ligand structure in the data set was regenerated using Corina. For each structure, protonation and tautomeric states were preserved from the original DUD set. For each target, all ligands and decoys were docked into the ligand-bound and *apo* protein structures given in the Astex non-native set.

Two protein structures were removed from the data set. The aldose reductase structure 1z8a was removed as residue LEU300 is disordered, and in the protein conformation used in the Astex non-native set, this side chain blocks part of the binding site. This was not an issue in the case of the ligand used from the Astex diverse set as it did not occupy the volume of the cavity blocked by LEU300. The second structure removed was an acetylcholine esterase structure, PDB code 2c5g. In this structure, the entrance to and part of the binding site is blocked by a crystallization reagent (trimethyl-(2-sulfanylethyl) azanium). Again, this structure does not present a problem in the Astex diverse set as the bound ligand did not occupy that region of the binding site.

For the assessment of global virtual screening performance we used the *receiver operator characteristic* (ROC) AUC (area under curve) metric.[65] Briefly, using a random compound selection one would expect to obtain an AUC of 0.5, whereas if all the actives were ranked better than the inactives one would get an AUC of 1.0. Although the AUC is a useful metric for evaluating the global performance of a virtual screen it does not capture vital information about early enrichment, which is of practical importance as only a small subset of the compounds from the virtual screen can be tested experimentally. For this reason also enrichment factors were calculated at the percentage of the database where all actives (EF-AA) of the specific target could be found. This measure relates the fraction of active ligands the docking method identifies in the top x% of the ranked database to the number of active ligands that would be retrieved by a random selection strategy. As the screening sets studied in this work contain 2−3% active ligands,

Table 2. Virtual Screening Test Set

		protein structures		DUD set		chemotypes	
target	PDB code	no. ligand-bound	no. apo	no. actives	no. decoys	no. actives	no. chemotypes
acetylcholine esterase	1gpk	20	0	105	3714	100	19
aldose reductase	1t40	31	2	26	918	26	14
cyclin-dependent kinase 2	1ke5	72	9	50	1778	47	32
dihydrofolate reductase	1s3v	9	1	201	7144	191	14
factor Xa	1lpz	34	0	142	5095	64	19
neuraminidase	1l7f	13	13	49	1744	49	7
p38 MAP kinase	1ywr	31	6	256	8386	137	20
phosphodiesterase 5A	1xoz	5	0	51	1809	26	22

dx.doi.org/10.1021/ci2005934 | J. Chem. Inf. Model. 2012, 52, 1262−1274

maximum enrichment factors of around 36 can be achieved, while a random selection strategy would be expected to result in an enrichment factor of 1.

We also investigated the effect of ensemble docking on chemotype enrichment. For this purpose, the chemotype-clustered version of the DUD data set[57] was used, which was generated from reduced graph representations of the active ligands.[66] Note that in the chemotype-clustered DUD set fewer active ligands were available for some targets due to the application of additional drug-likeness filters[66] (see Table 2), while the decoy set was the same as that in the original DUD set. For the assessment of chemotype enrichment performance, we applied the cluster-averaged versions of the ROC AUC and enrichment factor metrics,[56] i.e. the weight of each active ligand is inversely proportional to the size of the chemotype cluster it belongs to. The cluster-averaged enrichment factors were again determined at the percentage of the database where potentially all actives could be found.

Ensemble Scoring. There are several ways one can determine the scoring function value of a ligand, given an ensemble of protein structures. Assuming the individual scoring function values for each protein structure are available, common protocols select the best one or calculate an average value. When calculating an ensemble-averaged value usually either the arithmetic mean or a Boltzmann-weighted average is used.[47] Using an average scoring function value as the objective function value in a global optimization methodology has the disadvantage that it requires the evaluation of a ligand's poses in all protein structures of the ensemble, which can be very time-consuming for large ensembles. In contrast, if only the best scoring function value of a ligand across all ensemble members is of interest, efficient optimization strategies using the protein conformation as an independent variable of the optimization problem can be designed.[30,31,35] Although we have extended the genetic algorithm in GOLD to efficiently search ensembles in a similar manner, all results presented in this work are based on selecting the best scoring function value across all ensemble members for a given ligand.

Ensemble Enumeration. Given a set of n protein conformations, in total $2^n - 1$ ensembles containing at least one protein structure can be generated. For a specific subset of size k, there are exactly

$$\binom{n}{k}$$

ensembles. Exhaustively enumerating all ensembles for large values of n is computationally infeasible. For example, enumerating all possible ensembles for the 72 ligand-bound protein structures of the *cdk2* data set would result in $2^{72} - 1$ (around 4.7 thousand billion billion) combinations. Thus, where there are more than 500 000 possible combinations for a specific subset of size k, 500 000 nonredundant ensembles are enumerated by random sampling (see Figure 1). Applying this sampling methodology to the targets in this work resulted in between 31 (*pde5*) and 32.6 million (*cdk2*) different ensembles. Exhaustive enumeration of all possible ensembles was possible for the targets *ache*, *dhfr*, *na*, and *pde5*.

Ensemble pose prediction and virtual screening results were then generated by postprocessing the single-docking results. Given an ensemble selection generated by the methodology described above, for each ligand the pose with the highest scoring function value obtained across all protein structures in the ensemble was selected. In a pose prediction experiment, the rmsd of this pose compared to the experimentally observed ligand conformation was assessed. When multiple ligands were available for the same target, the average percentage of correctly docked ligands was calculated. Virtual screening was simulated for a given ensemble by first extracting a ranking order for the actives and decoys based on the ensemble member docking results. The ranking order was then used to calculate an AUC and enrichment factor. This led to sets of AUCs and enrichment factors for each set of ensembles of a given size k. For each of these sets, a maximum, minimum, and average AUC value and enrichment factor was determined. This allowed us to assess the average performance as well as the worst and best performing ensemble of size k.

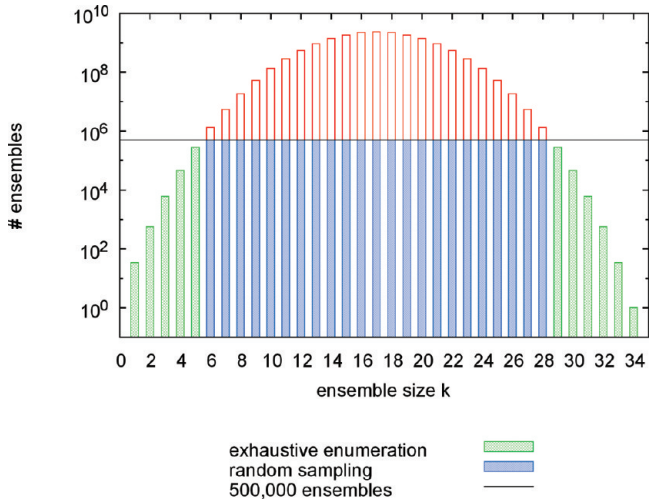


Figure 1. Illustration of the number of ensembles as a function of ensemble size k for $n = 34$ protein structures. For a given ensemble size k up to 500 000, ensembles are enumerated exhaustively (green bars), while in all other cases 500 000 random ensembles (blue bars) are generated by nonredundant random sampling.

Using these results, we tried to answer the question of whether ensemble docking performs better on average than the average of the n initial protein structures. This relates to the question of whether picking a single protein structure out of a set of n protein structures performs worse than an ensemble of size $k > 1$ picked at random.

Ligand Similarity. Part of this work deals with the influence of ligand similarity in the context of virtual screening performance. Pipeline Pilot[67] was used to calculate the Tanimoto similarities between all ligands of the DUD set and cocrystallized ligands of the ligand-bound protein structures based on ECFP6 fingerprints.[68]

RESULTS AND DISCUSSION

Pose Prediction. The pose prediction experiments were carried out using a subset of the Astex non-native set. For each protein target, each ligand was docked into all protein models. As we were interested in the non-native docking performance, the native protein model of each ligand was excluded. Poses with an rmsd better than 2.0 Å were considered to be correct. Ensemble docking results were simulated using the sequential docking results. All possible ensembles were enumerated and the worst, average, and best performing ensembles were compared for all possible ensemble sizes. Note that an ensemble size of one represents a classic single protein cross-docking experiment.

dx.doi.org/10.1021/ci2005934 | J. Chem. Inf. Model. 2012, 52, 1262−1274

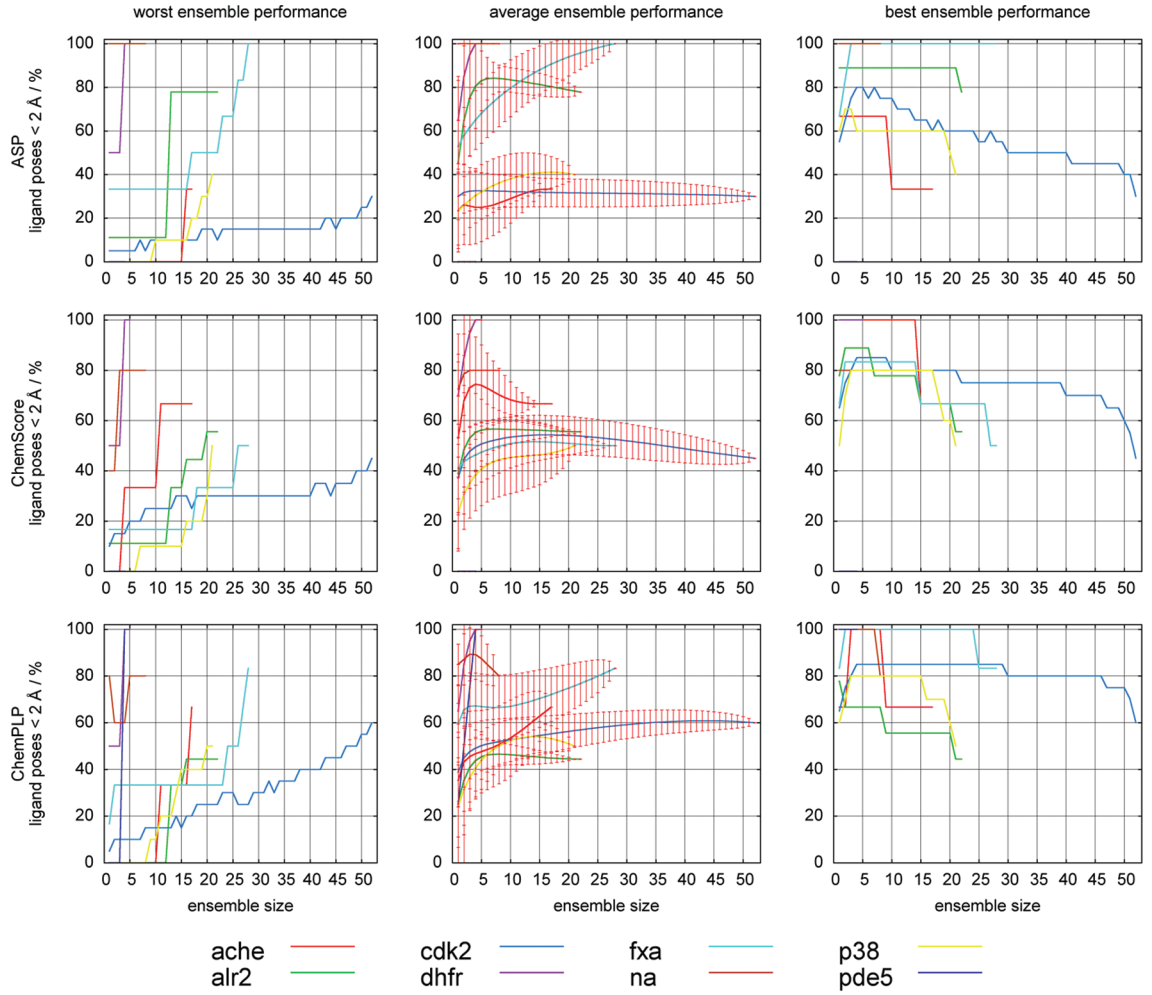


Figure 2. Non-native pose prediction results obtained for the eight targets and the three scoring functions. For each scoring function the plots from left to right show the performance of the worst, average, and best ensembles, respectively. Standard deviations are reported for the average performance.

Clear improvements in the pose prediction performance are observed as the number of protein models in the ensemble increase (see Figure 2). When looking at the worst performing ensembles, an improvement in the pose prediction performance can be observed when comparing the single protein results to ensemble results at the maximum number of protein structures available. With few exceptions, the average performance increases when using ensembles of protein structures. In many cases an increase of pose prediction performance with ensemble size can be observed. However, in some cases fewer than the maximum number of available structures on average results in the best pose prediction performance. Note that the worst, best, and average performance converge to the same value at the maximum ensemble size as there is only one ensemble of size *n*.

In general, the performance of the best performing ensembles decreases with increasing ensemble size. However, in many cases ensembles of size two or greater can be identified which perform better than the best single protein structure. Table 3 shows a subset of the cross-docking matrix obtained for docking six *fxa* ligands into all noncognate protein structures using the ChemPLP scoring function. Five out of the 28 protein structures used are able to reproduce five out of the six ligands correctly (shown in the upper part of the table). However, when adding a second protein structure, seven ensembles of size two (PDB codes 1mq5/2j34, **1nfu/2j34**, 1nfw/1nfy, 1nfw/2j34, **1nfy/2j34**, 1nfy/2uwo, and 2j2u/2j4i) are able to outperform the single protein structures and reproduce the conformation of all ligands correctly. Only two of the seven ensembles (shown in bold) consist of two of the best performing single protein structures, while the other five combine one of these with a poorer performing one. One of the six ligands (PDB code 2bok) clashes with the side chain conformation of GLN192 in protein structure 1mq5. Adding the second protein structure (PDB code 2j34) allows this ligand to be reproduced correctly, because the correctly predicted ligand structure in this protein has a higher fitness value and will thus be returned by the ensemble docking protocol. Although less common, an ensemble of size two or greater can in principle also perform less well than the worst single protein structure. This can be observed in *na* for ensembles of size two and three when using the ChemPLP scoring function (see lower left plot of Figure 2).

The chance of selecting an ensemble that performs well in pose prediction depends on two factors. First, a scoring function should be able to predict the correct ligand conformation in as many protein conformations as possible. Second, correctly docked solutions must be assigned better scoring function values than incorrect ones in order to be selected in an ensemble docking protocol. Given the single

dx.doi.org/10.1021/ci2005934 | *J. Chem. Inf. Model.* 2012, 52, 1262−1274

Table 3. Pose Prediction Results for Factor Xa Using the ChemPLP Scoring Function[a]

	Protein									
	1nfu		**1nfy**		**2g00**		**2j2u**		**2j34**	
ligand	RMSD	fitness	RMSD	fitness	RMSD	fitness	RMSD	fitness	RMSD	fitness
1ezq			1.8	111.2	1.8	112.2	1.9	91.7	5.1	90.0
1lpz	1.9	96.0	3.3	96.6	1.6	98.2	1.6	99.7	1.7	101.9
1mq6	1.8	97.0	1.9	85.9	1.4	99.2	1.7	96.6	1.7	96.0
1xka	1.5	101.3	1.9	97.9	1.7	107.1	1.6	97.4	1.4	98.8
2bok	5.8	66.6	1.6	69.8	3.6	71.5	2.3	70.6	1.5	68.1
2fzz	1.6	103.6	1.6	97.3	1.6	114.1	1.5	96.1	1.8	92.1
% correct	83.3		83.3		83.3		83.3		83.3	

	Protein							
	1mq5		1nfw		2j4i		2uwo	
ligand	RMSD	fitness	RMSD	fitness	RMSD	fitness	RMSD	fitness
1ezq	1.7	100.8	1.8	110.8	9.0	87.5	5.1	91.4
1lpz	9.9	86.0	1.9	97.5	3.4	99.2	1.7	96.7
1mq6	1.4	104.6	8.4	84.8	1.9	92.9	1.6	96.6
1xka	1.4	108.8	1.7	104.4	1.7	100.6	1.3	103.1
2bok	5.8	65.4	4.5	66.8	1.9	71.7	3.	68.0
2fzz	1.5	104.8	1.7	97.3	1.4	94.7	1.8	99.6
% correct	66.7		66.7		66.7		66.7	

[a]For each ligand and protein structure, the RMSD and fitness of the top-scoring ligand conformation are reported. The last row reports the average percentage of correctly predicted ligand conformations. Correct predictions, i.e. RMSD values lower than 2 Å, are color-coded in green while incorrect ones are shown in red. PDB codes of the best-performing single protein structures are highlighted in bold (upper part of the table).

Table 4. Neuraminidase Pose Prediction Results (Scoring Function ChemPLP)

ligand	1iny (EPI = 0.46)				1f8b (EPI = 0.97)			
rank	PDB	score	rmsd	correct	PDB	score	rmsd	correct
1	1f8c	75.49	4.24	×	1mwe	69.13	0.45	√
2	1f8e	75.47	0.82	√	2c4a	68.88	0.44	√
3	1mwe	75.17	0.84	√	1f8c	67.70	0.45	√
4	1f8d	73.12	0.80	√	2c4l	66.68	0.44	√
5	2c4l	72.45	4.30	×	1f8e	64.93	4.24	×
6	2qwj	72.19	1.04	√	2qwj	64.78	0.58	√
7	2c4a	71.98	4.36	×	1f8d	62.72	0.46	√
8	2qwi	69.57	1.10	√	2qwi	60.79	0.53	√

protein pose prediction results, we developed the *ensemble performance index* (EPI)[69] which captures both characteristics at the same time. Starting from a ranked list of single pose prediction results sorted by decreasing fitness values, the total number of correct solutions that would be obtained when enumerating all possible ensembles of at least size one were derived:

$$N = \sum_{r=1}^{n} I_r 2^{n-r} \tag{1}$$

In this equation, n is the number of protein structures and I_r is an indicator function which returns 1 if the solution at rank r is correct and 0 otherwise. When normalizing the total number of correct solutions by the total number of ensembles of size one or greater, $2^n - 1$, the following ensemble docking performance measure is derived

$$\mathrm{EPI} = \frac{N}{2^n - 1} \tag{2}$$

Like the AUC measure used in virtual screening calculations, this measure has a number of useful properties. The values of EPI range from 0 to 1, where 0 indicates that none and 1 that all of the solutions in the list of single protein pose prediction results were correct. Furthermore, it accounts for the ranking of correct and incorrect solutions in a sensible way. A value greater than 0.5 implies that the top-ranked solution is correct and thus increasing the ensemble size k up to n will necessarily converge to a correct solution. In contrast, if the top-ranked solution is wrong, the EPI value will be lower than 0.5 and the solution resulting from the ensemble of maximum size n will be wrong. Table 4 presents pose prediction results of two ligands docked to eight *na* protein structures using the ChemPLP scoring function. The ligand of PDB code 1iny has an EPI value of 0.46, which means 46% (i.e., 117) of all possible (i.e., $2^8 - 1$)

dx.doi.org/10.1021/ci2005934 | J. Chem. Inf. Model. 2012, 52, 1262−1274

ensembles, will result in a correct prediction. As the top-ranked solution over all protein structures has an rmsd higher than 2 Å, this is below 0.5. In contrast, the ligand of PDB code 1f8b gets a near-optimal EPI value of 0.97, with only one incorrectly predicted solution at a low rank. In this case 97% of all possible ensembles deliver a correct solution.

An overview of EPI values obtained for the individual targets can be found in Figure 3. The figure shows data for all scoring functions used. For each target the best, worst, and the average EPI values over all ligands are presented when docking them in the non-native protein conformations as described above. For *pde5*, the minimum, average, and maximum EPI values are the same as there is only one ligand. Scoring function ChemPLP is predicting the conformation of this ligand correctly in 8 out of 15 possible non-native protein ensembles, i.e. in 53% of the cases (EPI value of 0.53). Note that for a specific target the minimum and maximum EPI values observed across the three scoring functions do not necessarily correspond to the same ligand. Apart from target *dhfr*, the ensemble pose prediction performance is dependent on the scoring function used. When looking at the average EPI values, only for targets *dhfr*, *fxa*, and *na*, more than half of the ensembles result in a correct pose prediction across all three scoring functions. Overall, given the large standard deviations on the average EPI values it is hard to suggest which scoring function to use for which target.

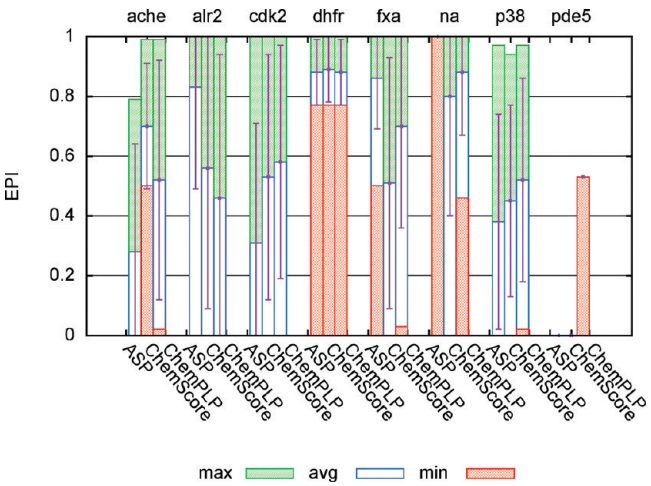


Figure 3. Minimum (red bars), maximum (green bars), and average (empty bars) EPI (ensemble performance index) values over all ligands listed per target in Table 1. For each of the eight targets the bars from left to right correspond to the results for ASP, ChemScore, and ChemPLP, respectively.

Virtual Screening. The virtual screening experiments were carried out on a subset of targets from the DUD data set. Actives and decoys were docked into protein models taken from the Astex non-native set. The ROC AUC metric and enrichment factors calculated at the percentage of the database where all actives could be found (EF-AA) were used to assess the success of the virtual screening experiments. Chemotype enrichments were also calculated using the chemotype-clustered version of the DUD data set. Exhaustive enumeration of all ensembles was carried out when computationally feasible; otherwise, 500 000 possible combinations per ensemble size k were selected at random. For each ensemble size k, the worst, average, and best performing ensembles were identified and compared. It is again worth noting that an ensemble of size one represents a classic single protein docking experiment.

The virtual screening performance of individual protein structures can vary dramatically (Figure 4). For example, for *p38* the worst performing structure (PDB code 2gfs) out of the 31 ligand-bound protein structures available in the data set only achieves an AUC value of around 0.4, while the best one reaches an AUC value of around 0.74 (PDB code 1w82). Similarly pronounced differences can be observed for targets *ache*, *alr2*, *cdk2*, and *fxa*. In general it is unknown, a priori, how a specific protein structure will perform in virtual screening and in the worst case the worst-performing structure available may be selected.

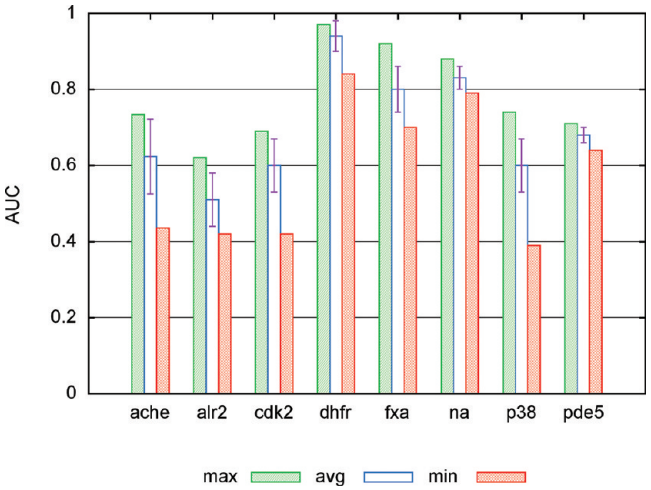


Figure 4. Statistics of the single protein structure screening results obtained for the ChemPLP scoring function. The actives and decoys were docked into all ligand-bound protein structures available for each target and AUC values were calculated. For each of the eight targets, the AUC value observed for the best and the worst protein structure are reported as well as the average AUC value over all protein structures (with standard deviations).

When using ligand-bound protein structures, trends similar to the pose prediction results can be observed (see Figure 5; enrichment factors can be found in the Supporting Information). With a few exceptions, the worst case performance improves with increasing ensemble size and the best case performance decreases. For some targets like *ache*, *cdk2* and *fxa* major improvements in the worst-case performance can be observed. While for single protein structures and small ensemble sizes AUC values below 0.5 are obtained for several targets and scoring functions, for large ensemble sizes AUC values of at least 0.5 are achieved across all scoring functions and targets. This consistent improvement is an important observation in the sense that ensemble docking might be seen as a risk minimization strategy.

The occurrence of exceptions, i.e. an ensemble giving a result worse than the worst protein structure (ChemScore for target *na*, ensemble size two) or better than the best single protein structure (ChemPLP and ASP for targets *ache*, *cdk2*, and *pde5*), can be explained by specific ranking patterns. Table 5 illustrates a hypothetical virtual screening scenario considering two protein structures and a database consisting of two active ligands (L1 and L2) and one decoy (D). The ranking obtained for protein structure 1 is [L1, D, L2] and [L2, D, L1] in protein structure 2. For each ranking the corresponding scoring

dx.doi.org/10.1021/ci2005934 | J. Chem. Inf. Model. 2012, 52, 1262−1274

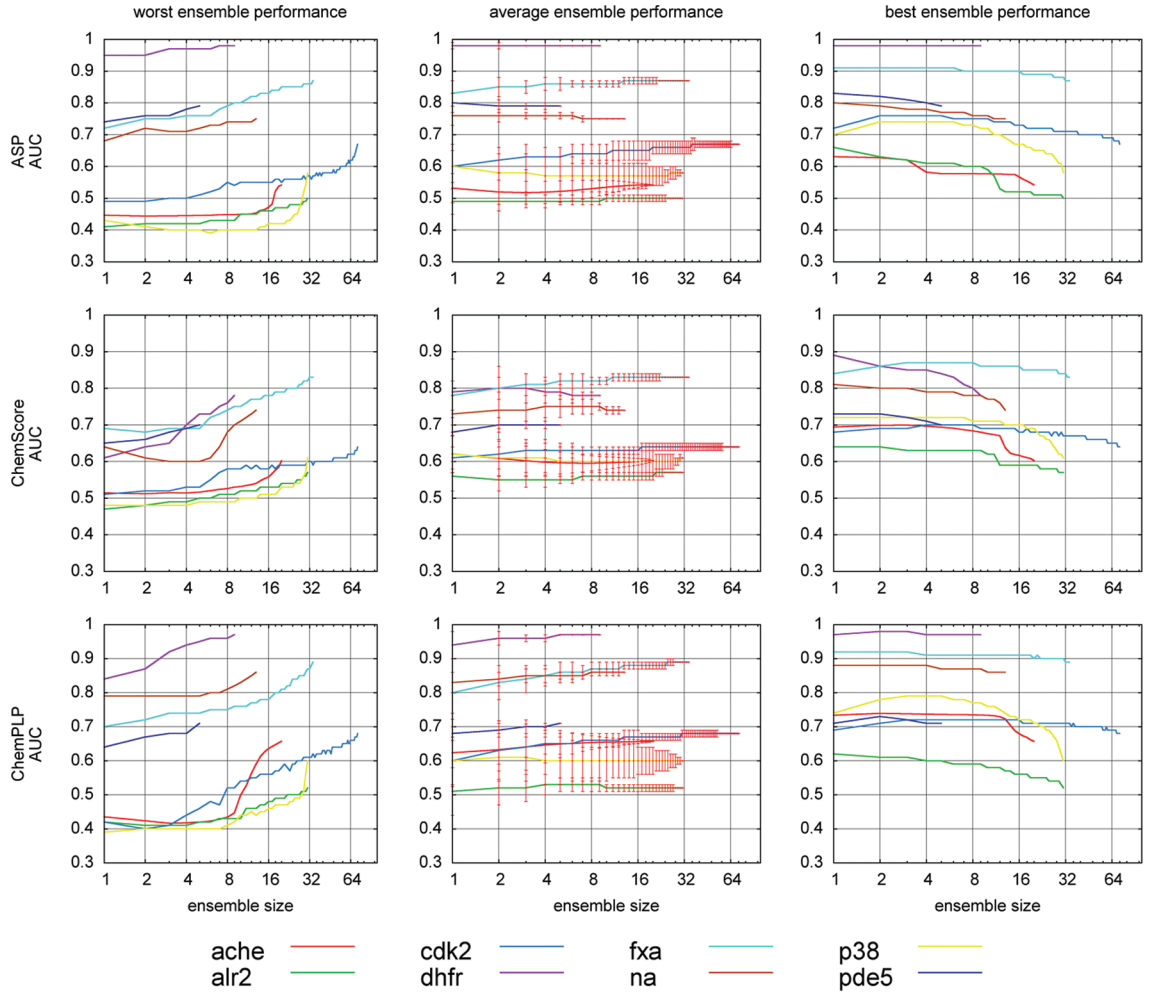


Figure 5. Virtual screening results (AUC). The columns from left to right correspond to the performance of the worst ensemble, the average over all ensembles (standard deviations are reported), and the performance of the best ensemble, respectively. Note the logarithmic scale of the *x*-axis.

Table 5. Improving upon the Best Single Protein Structure Performance

protein 1 (AUC 0.5)		protein 2 (AUC 0.5)		ensemble (AUC 1.0)	
rank	score	rank	score	rank	score
L1	**70**	**L2**	**60**	**L1**	70
D	**50**	**D**	45	**L2**	60
L2	40	**L1**	30	**D**	50

function values are also presented. When applying an ensemble docking protocol and selecting the highest scoring value for each ligand, the initial single protein structure rankings [L1, D, L2] and [L2, D, L1] can be improved to [L1, L2, D]. However, the same scenario can also lead to a decrease in performance when exchanging ligands and decoys.

When looking at the average performance, changes in performance with increasing ensemble size are less pronounced. Nevertheless, virtual screening against some targets like *dhfr*, *fxa*, and *na* significantly benefits from ensemble docking.

Chemotype Enrichment. Let us look at the chemotype enrichment in the context of ensemble docking. The plain AUC values obtained for all targets, scoring functions, and ensemble sizes were compared to their respective cluster-averaged values (cluster-averaged enrichment factors can be found in the Supporting Information) and are presented in Figure 6. Note that the enrichment factors differ from the cluster-averaged chemotype enrichment factors reported in Figure 5 due to the different number of active ligands considered in the chemotype-clustered version of the DUD data set (see Table 2). While for many targets only minor differences between the plain and cluster-averaged AUC values can be observed, some targets show major discrepancies. For *ache* and *dhfr*, the AUC values are much higher than the cluster-averaged values, while for *na* the chemotype enrichment is higher. In agreement with another study on chemotype enrichment using the same data set and the DOCK program,[56] we also find that the cluster-averaged AUC values for *fxa* are much higher than the plain AUC values. Compared to the virtual screening results presented before, similar trends can be observed with increasing ensemble size.

Sampling Problem. Most software packages provide virtual screening settings which spend only a fraction of the standard docking time per ligand to allow for the screening of large libraries. However, this increase in speed comes at the cost of sampling accuracy. When trying to extract structure-based rules for the selection of well-performing protein ensembles, these settings will influence the outcome and in the worst case wrong conclusions may be drawn. In order to quantify the effect of sampling accuracy on ensemble docking performance, we performed the same kind of analyses as presented above when using between 1 and 20 genetic

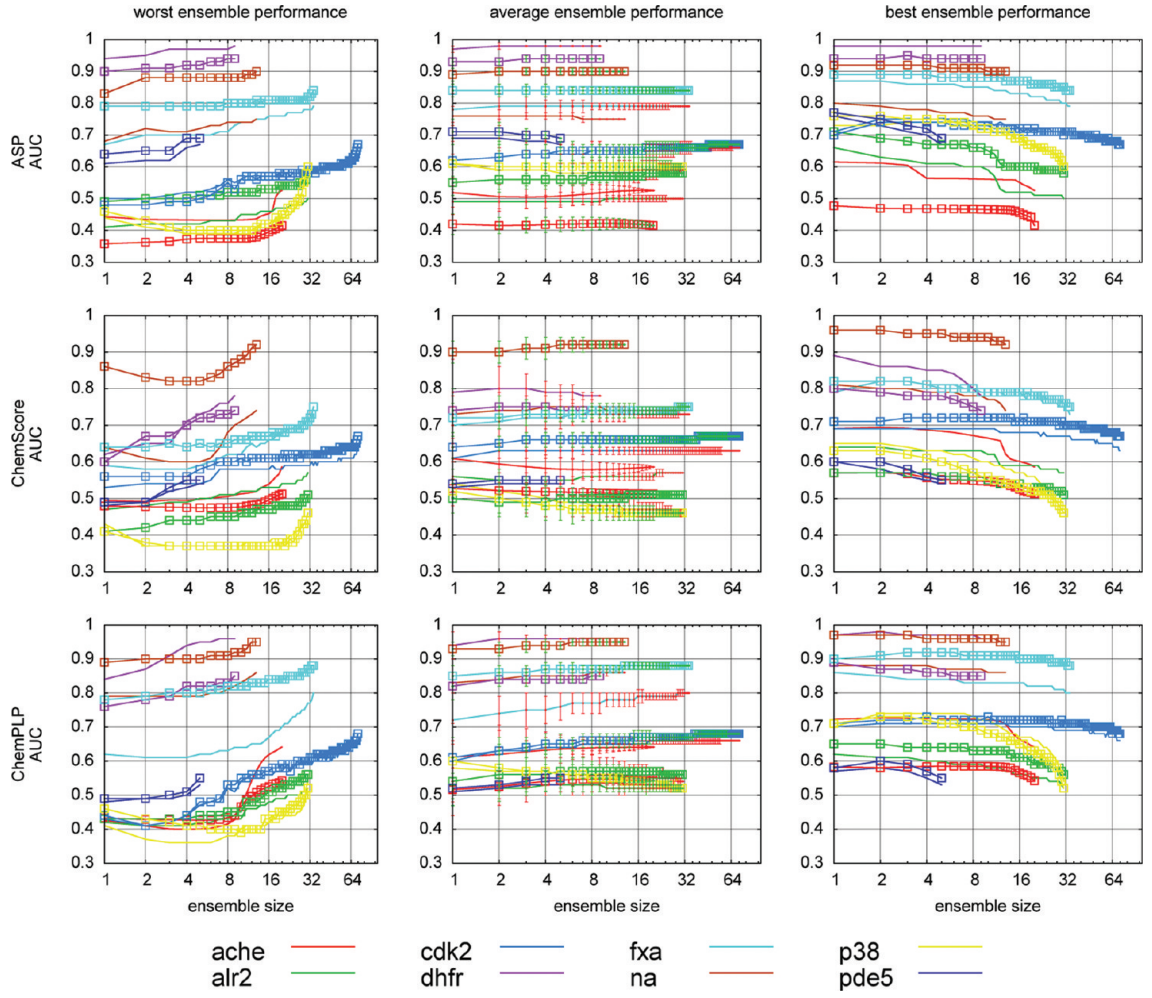


Figure 6. Comparison of enrichment factors and cluster-averaged chemotype enrichment factors. Cluster-averaged AUC values calculated using the chemotype classification are marked with rectangles. Note the logarithmic scale of the *x*-axis.

algorithm (GA) runs per ligand. We restricted the targets to *ache*, *dhfr*, *na*, and *pde5*, those where all possible ensembles can be enumerated.

Figure 7 presents the virtual screening results as a function of the ensemble size for different numbers of GA runs. As expected, the ensemble docking results are in some cases crucially dependent on the sampling accuracy of the docking algorithm. While for *ache* a better screening performance can be obtained when increasing the sampling accuracy, this is not necessarily the case for all targets. For *dhfr* the fastest search setting, i.e. one GA run, performs worst in the single protein structure case but performs slightly better than the longest search settings when looking at the multiple protein structure results. For the *na* and *pde5* data sets, the fastest search setting performs the best and the most time-consuming one the worst across all ensemble sizes when looking at the AUC values. This suggests that in these data sets actives can on average be docked "easily", i.e. reach on average higher scores, than decoys, when given a very short sampling time. With respect to enrichment factors, there is no clear trend observable, but the ensemble docking performance is again affected by the sampling accuracy.

These experiments also revealed that the best-performing ensembles for a given ensemble size are dependent on the sampling time. For example, the best performing ensembles of size 3 using 1 GA run may not be the best ones when using 20 GA runs. This should be kept in mind when training sets are used to derive optimum ensembles to be used in a large-scale virtual screening campaign or when computationally deriving structure-specific binding energy offsets.[41]

Other Sources of Protein Structures. The same analysis has also been carried out for *apo* X-ray structures and a ligand-bound NMR structure (25 protein conformations, PDB code 1YHO) of *dhfr*. In agreement with observations of other researchers,[37,70] we found that using ensembles of *apo* crystal structures usually resulted in an inferior pose prediction and virtual screening performance compared to ligand-bound ones. Nevertheless similar trends were apparent, i.e. the worst-case performance is usually improved when using multiple *apo* structures. Results for the NMR-derived *dhfr* ensemble and the *apo* protein structures are presented in the Supporting Information.

Impact of Ligand Similarity. Other studies have shown that the similarity of a ligand to the cocrystallized ligand of the protein structure in which it is docked has an impact on the pose prediction success rate.[22,24,71] In order to study the effect of ligand similarity on virtual screening results in the context of ensemble docking, for targets *ache*, *dhfr*, *na*, and *pde5*, we removed all ligands and decoys which had a Tanimoto similarity higher than 0.4 (ECFP6 fingerprints) to any of the cocrystallized ligands of their respective sets of ligand-bound

dx.doi.org/10.1021/ci2005934 | J. Chem. Inf. Model. 2012, 52, 1262−1274

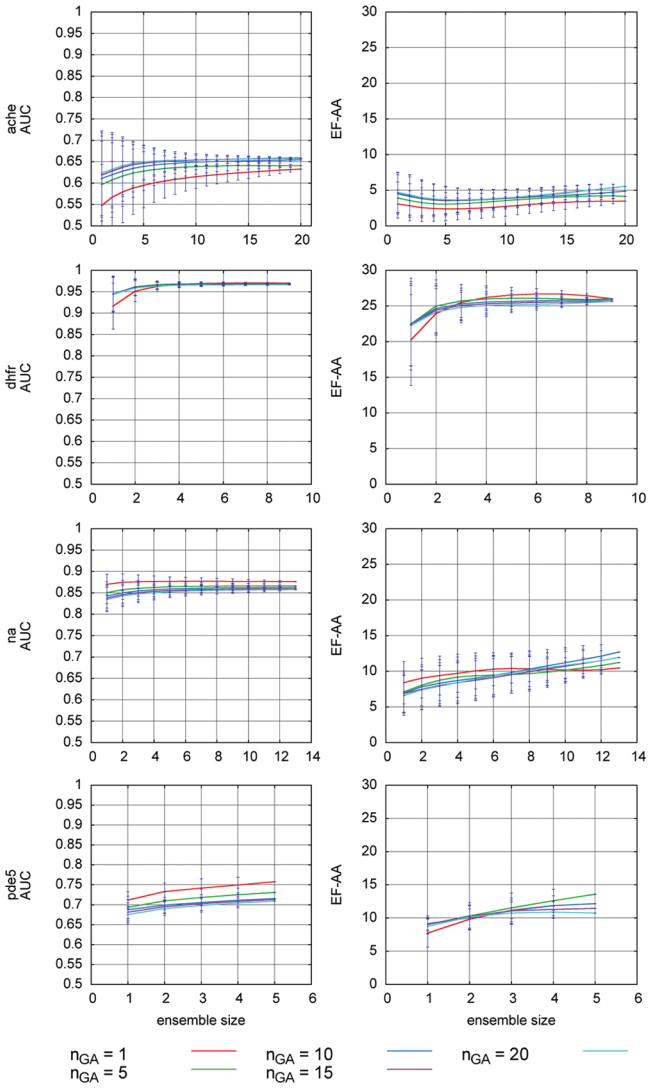


Figure 7. Ensemble virtual screening results as a function of the ensemble size for different numbers of GA runs. For each GA setting, the average AUC value and enrichment factor (with standard deviations) over all possible ensembles are presented, respectively.

protein structures. Table 6 summarizes the percentage of ligands and decoys left after this filter. In all cases, the ligand

Table 6. Number of Actives and Decoys Left after the Ligand Similarity Filtering

Target	no. actives (% of original)	no. decoys (% of original)
ache	78 (74.29%)	3713 (99.97%)
dhfr	95 (47.26%)	7143 (99.99%)
na	15 (30.61%)	1715 (98.34%)
pde5	43 (84.31%)	1809 (100.00%)

sets contained more similar structures than the decoy sets. For targets *dhfr* and *na* only 30−50% of the original active data set is left, while the decoy sets are nearly unaltered.

This also implies that fingerprint based methods already perform very well for these targets and the benefit of using a three-dimensional search methodology like docking is not obvious. Figure 8 presents the ensemble virtual screening results for the four targets when using scoring function ChemPLP. While for *ache*, *dhfr*, and *pde5*, only a small

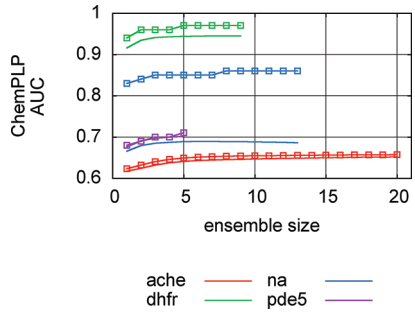


Figure 8. Comparison of the average AUC values (scoring function ChemPLP) over all ensembles of the respective size for the targets *ache*, *dhfr*, *na*, and *pde5* when using the ligand-similarity filtered and original data sets (marked with rectangles). For clarity, standard deviations for the average values are omitted.

decrease in virtual screening performance can be observed for the filtered data set, the AUC values for *na* are significantly lower compared to the original data set (around 0.65 versus around 0.85). These results imply that the use of ligand similarity information might be helpful in the context of ensemble structure selection, as will be shown later.

Scoring Function Dependence. As already shown above, the three scoring functions can perform quite differently for the same target. We looked, whether for a specific ensemble size, the best performing ensembles for each scoring function are identical. Due to the exhaustive enumeration of all possible ensembles for four of our targets, the maximum AUC values and enrichment factors theoretically achievable for each ensemble size are known. So, for each ensemble, we compared the performance of each individual scoring function with the known optimum value. Figure 9 shows the enrichment factors

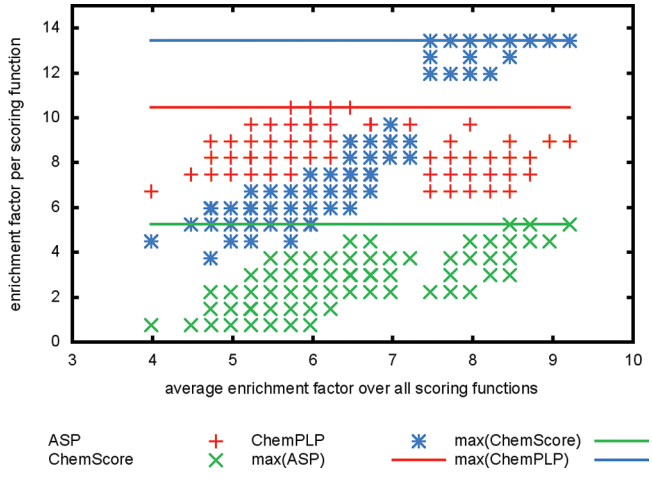


Figure 9. Comparison of neuraminidase enrichment factors (EF-AA) across all scoring functions for all ensembles of size three. For each scoring function, the maximum enrichment factor observed for this ensemble size is highlighted. The ensembles are sorted according to the average enrichment factor over the three scoring functions.

obtained for all ensembles of size three across all scoring functions for *na*. The ensembles are sorted according to the average enrichment factors across all scoring functions. Additionally, the maximum enrichment factor observed for each scoring function across all ensembles of size three are marked by bold lines for each scoring function. In theory, optimum ensembles, i.e. ensembles where all scoring functions

dx.doi.org/10.1021/ci2005934 | J. Chem. Inf. Model. 2012, 52, 1262−1274

reach their maximum enrichment factor, should be observable on the right-hand side of the plot. However, in the example presented, there are three ensembles where ChemScore and ChemPLP perform optimally, but ASP performs suboptimally. In general, it is hard to find ensembles that perform optimally for all three scoring functions. This can be explained by the fact that the scoring functions reward certain types of contributions, like hydrogen bonding or van der Waals interactions, differently. As a consequence, selecting well performing ensembles based on protein structure related features, like binding site diversity etc., and not taking into account scoring function dependent information, is a challenging task.

Ensemble Selection Protocols. GOLD has been tuned to search up to 15 protein structures concurrently using an efficient search method. In practice, however, a modeler may be faced with the opportunity to use more than 15 protein structures. A methodology for making an informed decision on which protein models to include in the ensemble would therefore be desirable. Here we reassessed the performance of published selection methodologies by comparing their efficiency with the results obtained from our large-scale analysis for all targets where more than ten ligand-bound protein structures were available. Note that our test sets and scoring functions are different to the ones used in the published studies and different results might therefore be expected. Protocols for the selection of protein ensembles are based on (i) the size of the cocrystallized ligand of a ligand-bound protein structure, (ii) the fingerprint similarity of database ligands to cocrystallized ligands in the ligand-bound protein structures, and (iii) the mean ligand score of a subset of the docked database.

Protocol i follows the study of Rueda et al.[37] where all available ligand-bound protein structures are sorted according to a decreasing number of heavy atoms of their cocrystallized ligands. Protocol ii is inspired by studies[22,24,71] using the fingerprint similarity of the ligands to be docked and the cocrystallized ligands of the ligand-bound structures to select the protein structures for docking. Both studies report improved pose prediction success rates when the ligand similarity information is taken into account. The last protocol iii has been applied in an ensemble docking study on p38 MAP kinase where the ensembles have been selected based on the mean ligand score of the top 1% of the docked database.[38]

Before discussing the results obtained when applying the protocols, a few advantages and disadvantages of each approach should be highlighted. The first two approaches rely on protein structure related data only, but do not take into account any scoring function specific information. While the number of ligand atoms in protocol i could be replaced by a binding site volume descriptor, and thus also be used with *apo* protein conformations, the ligand similarity-based protocol ii needs a ligand-bound protein structure. In addition, for protocol ii, the situation can arise that the ligands of only a few protein structures are similar to any database ligands used for the screening. This is, for example, the case for target *ache*, where database ligands are only similar to the cocrystallized ligands in three out of 20 protein structures. In this situation, either only three protein structures may be used or a few arbitrarily chosen structures could be added. As shown before, the selection of optimum ensembles is scoring function dependent, and thus, it is expected that these two protocols are affected by this problem, i.e. the ensembles selected, although potentially optimal for one scoring function, will possibly be suboptimal for another one. Protocol iii exclusively uses information from single protein structure docking calculations and therefore takes scoring function specific information into account. Additionally, it can be applied to *apo* protein structures or structures extracted from molecular dynamics simulations. A big disadvantage of this approach is that all single protein structure screening results need to be available, which is very time-consuming.

In general, across all targets, the protocols are capable of selecting ensembles which perform better than the worst ones identified by exhaustive enumeration. At the same time, they are not able to perform as well as the best known ensemble for respective particular ensemble size and, depending on the scoring function and target, the protocols can perform worse than the average performance expected from a random ensemble selection process. The AUC values obtained as a function of the ensemble size when applying these ensemble selection protocols are presented in the Supporting Information.

Protocol i using the number of ligand heavy atoms in the ligand-bound protein structures for the ensemble selection shows a highly variable performance. It performs reasonably well for only a few very specific target and scoring function combinations like *cdk2* (ASP and ChemPLP), *fxa* (ChemScore), *na* (ChemPLP), and *p38* (ChemPLP).

The ligand similarity based protocol ii on average performs best. This is consistent with observations of Tuccinardi et al.[71] who identified GOLD/ChemPLP combined with a ligand similarity based protein structure selection as one of the best performing protocols when used in cross-docking experiments of protein kinase structures.

Protocol iii usually selects ensembles which result in AUC values similar to the average ensemble performance as determined in the exhaustive enumeration experiments. It performs very well in combination with ASP and ChemPLP for *cdk2* and *na* but returns a suboptimal selection for *fxa* when used with the same scoring functions.

For most targets there is a large gap between the performance observed for the ensembles selected prospectively using the protocols presented and the performance of the best ensembles identified in the exhaustive enumeration experiments. Taking these results into account, we believe that the development of optimally performing ensemble selection protocols is a highly challenging and unsolved task.

CONCLUSIONS

In this work, we have presented a comprehensive study of the influence of ensemble size on virtual screening and pose prediction. For three scoring functions available in GOLD, ensembles for eight targets were enumerated exhaustively, where computationally feasible. In almost all cases, using ensembles of proteins performs better than the worst single protein structure. As the virtual screening performance of a single protein structure is unknown, a priori, the use of multiple protein structures in an ensemble docking protocol minimizes the risk of giving the worst virtual screening performance possible. Furthermore, in many cases, ensemble docking results in an improvement over the average of the single protein structure results. For some targets, ensembles that outperformed the best single protein structure could be identified. Similar results were also observed when looking at chemotype enrichment.

We identify several key factors affecting ensemble docking performance. Like single protein structure docking, the

performance of ensemble docking is scoring function and target dependent. Additionally, we see that a structural ensemble delivering an optimum performance for one scoring function does not necessarily perform well for a different scoring function. This implies that in theory optimal ensemble selection protocols need to take scoring function dependent information into account in order to circumvent this problem. We also see that the sampling accuracy of a docking approach has a crucial impact on the outcome of ensemble docking results. Large errors are introduced when using virtual screening settings giving insufficient sampling. In the worst case this may result in a suboptimal ensemble selection. In agreement with other authors[22,71] who observed that a high similarity between the ligand to be docked and the ligand cocrystallized with the protein it is docked to, usually results in higher cross-docking success rates, we make similar observations in the context of virtual screening and ensemble selection. Finally, we assessed the performance of several ensemble selection protocols and compared the results to the ones obtained from the exhaustive enumeration experiments. Usually, the ligand-similarity based protocol performs best but is still much worse than the best known ensembles for the respective targets. The development of ensemble selection protocols able to select optimally performing ensembles for different scoring functions and targets will be one of the major challenges in the future.

The internal energy of the protein structures was neglected in our studies, but we would expect it to have a large impact on the ensemble docking performance. While a prospective estimation of feasible values is extremely difficult, as different scoring functions have different scoring ranges, a retrospective fitting of optimum values based on a training set might be feasible.[41] Ultimately, in order to improve ranking performance, new ensemble scoring functions which identify, for a given ligand, the best fitting protein structure in the ensemble as the highest scoring must be developed. Due to the exhaustive nature of our experiments, we identified a large set of well-performing structural ensembles. Research is ongoing to characterize these ensembles and to identify features which might allow for a prospective identification of near-optimal ensembles. Our extensions to GOLD now allow us to search ensembles time-efficiently using tailored search methodologies.

■ ASSOCIATED CONTENT

Supporting Information

Enrichment factor plots for ligand-bound protein structures, AUC and enrichment factor plots for *apo* protein structures, comparison of enrichment factors and cluster-averaged chemotype enrichment factors, virtual screening results for an NMR-derived protein ensemble, and comparison of ensemble selection protocols and ligand similarities. This material can be downloaded free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +44-1223763923. Fax: +44-1223336033. E-mail: korb@ccdc.cam.ac.uk.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Dr. Colin Groom for carefully reading the manuscript and helpful discussions. O.K. was funded through a fellowship within the Postdoc-Programme of the German Academic Exchange Service (DAAD). This work was performed using the Darwin Supercomputer of the University of Cambridge High Performance Computing Service (http://www.hpc.cam.ac.uk/), provided by Dell Inc. using Strategic Research Infrastructure Funding from the Higher Education Funding Council for England.

■ REFERENCES

(1) Cozzini, P.; Kellogg, G. E.; Spyrakis, F.; Abraham, D. J.; Costantino, G.; Emerson, A.; Fanelli, F.; Gohlke, H.; Kuhn, L. A.; Morris, G. M.; Orozco, M.; Pertinhez, T. A.; Rizzi, M.; Sotriffer, C. A. Target Flexibility: An Emerging Consideration in Drug Discovery and Design. *J. Med. Chem.* **2008**, *51*, 6237−6255.

(2) Teague, S. J. Implications of protein flexibility for drug discovery. *Nat. Rev. Drug Discovery* **2003**, *2*, 527−541.

(3) Teodoro, M. L.; Kavraki, L. E. Conformational flexibility models for the receptor in structure based drug design. *Curr. Pharm. Des.* **2003**, *9*, 1635−1648.

(4) Henzler, A. M.; Rarey, M. In Pursuit of Fully Flexible Protein-Ligand Docking: Modeling the Bilateral Mechanism of Binding. *Mol. Inf.* **2010**, *29*, 164−173.

(5) B-Rao, C.; Subramanian, J.; Sharma, S. D. Managing protein flexibility in docking and its applications. *Drug Discovery Today* **2009**, *14*, 394−400.

(6) Koshland, D. E. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc. Natl. Acad. Sci. U.S.A* **1958**, *44*, 98−104.

(7) Gunasekaran, K.; Nussinov, R. How different are structurally flexible and rigid binding sites? Sequence and structural features discriminating proteins that do and do not undergo conformational change upon ligand binding. *J. Mol. Biol.* **2007**, *365*, 257−273.

(8) Gutteridge, A.; Thornton, J. Conformational changes observed in enzyme crystal structures upon substrate binding. *J. Mol. Biol.* **2005**, *346*, 21−28.

(9) Ferrari, A. M.; Wei, B. Q.; Costantino, L.; Shoichet, B. K. Soft Docking and Multiple Receptor Conformations in Virtual Screening. *J. Med. Chem.* **2004**, *47*, 5076−5084.

(10) Leach, A. R. Ligand docking to proteins with discrete side-chain flexibility. *J. Mol. Biol.* **1994**, *235*, 345−356.

(11) Hartmann, C.; Antes, I.; Lengauer, T. Docking and scoring with alternative side-chain conformations. *Proteins: Struct., Funct., Bioinf.* **2009**, *74*, 712−726.

(12) Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. FDS: flexible ligand and receptor docking with a continuum solvent model and soft-core energy function. *J. Comput. Chem.* **2003**, *24*, 1637−1656.

(13) Kokh, D. B.; Wenzel, W. Flexible Side Chain Models Improve Enrichment Rates in In Silico Screening. *J. Med. Chem.* **2008**, *51*, 5919−5931.

(14) Frimurer, T. M.; Peters, G. H.; Iversen, L. F.; Andersen, H. S.; Møller, N. P. H; Olsen, O. H. Ligand-induced conformational changes: improved predictions of ligand binding conformations and affinities. *Biophys. J.* **2003**, *84*, 2273−2281.

(15) Zhao, Y.; Sanner, M. Protein−ligand docking with multiple flexible side chains. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 673−679.

(16) Alberts, I. L.; Todorov, N. P.; Dean, P. M. Receptor Flexibility in de Novo Ligand Design and Docking. *J. Med. Chem.* **2005**, *48*, 6585−6596.

(17) Najmanovich, R.; Kuttner, J.; Sobolev, V.; Edelman, M. Side-chain flexibility in proteins upon ligand binding. *Proteins: Struct., Funct., Genet.* **2000**, *39*, 261−268.

(18) Sandak, B.; Wolfson, H. J.; Nussinov, R. Flexible docking allowing induced fit in proteins: Insights from an open to closed conformational isomers. *Proteins: Struct., Funct., Genet.* **1998**, *32*, 159−174.

(19) Koska, J.; Spassov, V. Z.; Maynard, A. J.; Yan, L.; Austin, N.; Flook, P. K.; Venkatachalam, C. M. Fully Automated Molecular Mechanics Based Induced Fit Protein−Ligand Docking Method. *J. Chem. Inf. Model.* **2008**, *48*, 1965−1973.
(20) Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. Novel Procedure for Modeling Ligand/Receptor Induced Fit Effects. *J. Med. Chem.* **2006**, *49*, 534−553.
(21) Bottegoni, G.; Kufareva, I.; Totrov, M.; Abagyan, R. A new method for ligand docking to flexible receptors by dual alanine scanning and refinement (SCARE). *J. Comput.-Aided Mol. Des.* **2008**, *22*, 311−325.
(22) Verdonk, M. L.; Mortenson, P. N.; Hall, R. J.; Hartshorn, M. J.; Murray, C. W. Protein-ligand docking against non-native protein conformers. *J. Chem. Inf. Model.* **2008**, *48*, 2214−2225.
(23) Erickson, J. A.; Jalaie, M.; Robertson, D. H.; Lewis, R. A.; Vieth, M. Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. *J. Med. Chem.* **2004**, *47*, 45−55.
(24) Sutherland, J. J.; Nandigam, R. K.; Erickson, J. A.; Vieth, M. Lessons in Molecular Recognition. 2. Assessing and Improving Cross-Docking Accuracy. *J. Chem. Inf. Model.* **2007**, *47*, 2293−2302.
(25) Corbeil, C. R.; Moitessier, N. Docking Ligands into Flexible and Solvated Macromolecules. 3. Impact of Input Ligand Conformation, Protein Flexibility, and Water Molecules on the Accuracy of Docking Programs. *J. Chem. Inf. Model.* **2009**, *49*, 997−1009.
(26) Jain, A. Effects of protein conformation in docking: improved pose prediction through protein pocket adaptation. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 355−374.
(27) Birch, L.; Murray, C. W.; Hartshorn, M. J.; Tickle, I. J.; Verdonk, M. L. Sensitivity of molecular docking to induced fit effects in influenza virus neuraminidase. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 855−869.
(28) Murray, C. W.; Baxter, C. A.; Frenkel, A. D. The sensitivity of the results of molecular docking to induced fit effects: application to thrombin, thermolysin and neuraminidase. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 547−562.
(29) Rueda, M.; Bottegoni, G.; Abagyan, R. Consistent Improvement of Cross-Docking Results Using Binding Site Ensembles Generated with Elastic Network Normal Modes. *J. Chem. Inf. Model.* **2009**, *49*, 716−725.
(30) Bottegoni, G.; Kufareva, I.; Totrov, M.; Abagyan, R. Four-Dimensional Docking A Fast and Accurate Account of Discrete Receptor Flexibility in Ligand Docking. *J. Med. Chem.* **2009**, *52*, 397−406.
(31) Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. FlexE: efficient molecular docking considering protein structure variations. *J. Mol. Biol.* **2001**, *308*, 377−395.
(32) Nabuurs, S. B.; Wagener, M.; de Vlieg, J. A Flexible Approach to Induced Fit Docking. *J. Med. Chem.* **2007**, *50*, 6507−6518.
(33) Wei, B. Q.; Weaver, L. H.; Ferrari, A. M.; Matthews, B. W.; Shoichet, B. K. Testing a flexible-receptor docking algorithm in a model binding site. *J. Mol. Biol.* **2004**, *337*, 1161−1182.
(34) Knegtel, R. M. A.; Kuntz, I. D.; Oshiro, C. M. Molecular docking to ensembles of protein structures. *J. Mol. Biol.* **1997**, *266*, 424−440.
(35) Huang, S.-Y.; Zou, X. Ensemble docking of multiple protein structures: Considering protein structural variations in molecular docking. *Proteins: Struct., Funct., Bioinf.* **2007**, *66*, 399−421.
(36) Craig, I. R.; Essex, J. W.; Spiegel, K. Ensemble Docking into Multiple Crystallographically Derived Protein Structures: An Evaluation Based on the Statistical Analysis of Enrichments. *J. Chem. Inf. Model.* **2010**, *50*, 511−524.
(37) Rueda, M.; Bottegoni, G.; Abagyan, R. Recipes for the Selection of Experimental Protein Conformations for Virtual Screening. *J. Chem. Inf. Model.* **2010**, *50*, 186−193.
(38) Rao, S.; Sanschagrin, P.; Greenwood, J.; Repasky, M.; Sherman, W.; Farid, R. Improving database enrichment through ensemble docking. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 621−627.
(39) Barril, X.; Morley, S. D. Unveiling the Full Potential of Flexible Receptor Docking Using Multiple Crystallographic Structures. *J. Med. Chem.* **2005**, *48*, 4432−4443.
(40) Limongelli, V.; Marinelli, L.; Cosconati, S.; Braun, H. A.; Schmidt, B.; Novellino, E. Ensemble-Docking Approach on BACE-1: Pharmacophore Perception and Guidelines for Drug Design. *ChemMedChem* **2007**, *2*, 667−678.
(41) Park, S.-J.; Kufareva, I.; Abagyan, R. Improved docking, screening and selectivity prediction for small molecule nuclear receptor modulators using conformational ensembles. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 459−471.
(42) Damm, K. L.; Carlson, H. A. Exploring Experimental Sources of Multiple Protein Conformations in Structure-Based Drug Design. *J. Am. Chem. Soc.* **2007**, *129*, 8225−8235.
(43) Hritz, J.; de Ruiter, A.; Oostenbrink, C. Impact of Plasticity and Flexibility on Docking Results for Cytochrome P450 2D6: A Combined Approach of Molecular Dynamics and Ligand Docking. *J. Med. Chem.* **2008**, *51*, 7469−7477.
(44) Lin, J.-H.; Perryman, A. L.; Schames, J. R.; McCammon, J. A. Computational Drug Design Accommodating Receptor Flexibility: The Relaxed Complex Scheme. *J. Am. Chem. Soc.* **2002**, *124*, 5632−5633.
(45) Bolstad, E. S. D.; Anderson, A. C. In pursuit of virtual lead optimization: Pruning ensembles of receptor structures for increased efficiency and accuracy during docking. *Proteins: Struct., Funct., Bioinf.* **2009**, *75*, 62−74.
(46) Bolstad, E. S. D.; Anderson, A. C. In pursuit of virtual lead optimization: The role of the receptor structure and ensembles in accurate docking. *Proteins: Struct., Funct., Bioinf.* **2008**, *73*, 566−580.
(47) Paulsen, J. L.; Anderson, A. C. Scoring Ensembles of Docked Protein:Ligand Interactions for Virtual Lead Optimization. *J. Chem. Inf. Model.* **2009**, *49*, 2813−2819.
(48) Park, I.-H.; Li, C. Dynamic Ligand-Induced-Fit Simulation via Enhanced Conformational Samplings and Ensemble Dockings: A Survivin Example. *J. Phys. Chem. B* **2010**, *114*, 5144−5153.
(49) Cheng, L. S.; Amaro, R. E.; Xu, D.; Li, W. W.; Arzberger, P. W.; McCammon, J. A. Ensemble-Based Virtual Screening Reveals Potential Novel Antiviral Compounds for Avian Influenza Neuraminidase. *J. Med. Chem.* **2008**, *51*, 3878−3894.
(50) Wong, C. F.; Kua, J.; Zhang, Y.; Straatsma, T. P.; McCammon, J. A. Molecular docking of balanol to dynamics snapshots of protein kinase A. *Proteins: Struct., Funct., Bioinf.* **2005**, *61*, 850−858.
(51) Yoon, S.; Welsh, W. J. Identification of a Minimal Subset of Receptor Conformations for Improved Multiple Conformation Docking and Two-Step Scoring. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 88−96.
(52) Cavasotto, C. N.; Kovacs, J. A.; Abagyan, R. A. Representing Receptor Flexibility in Ligand Docking through Relevant Normal Modes. *J. Am. Chem. Soc.* **2005**, *127*, 9632−9640.
(53) Novoa, E. M.; Pouplana, L. R.; de; Barril, X.; Orozco, M. Ensemble Docking from Homology Models. *J. Chem. Theory Comput.* **2010**, *6*, 2547−2557.
(54) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved protein-ligand docking using GOLD. *Proteins: Struct., Funct., Genet.* **2003**, *52*, 609−623.
(55) Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, *245*, 43−53.
(56) Mackey, M. D.; Melville, J. L. Better than Random? The Chemotype Enrichment Problem. *J. Chem. Inf. Model.* **2009**, *49*, 1154−1162.
(57) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.
(58) Mooij, W. T. M.; Verdonk, M. L. General and targeted statistical potentials for protein-ligand interactions. *Proteins: Struct., Funct., Bioinf.* **2005**, *61*, 272−287.
(59) Murray, C. W.; Auton, T. R.; Eldridge, M. D. Empirical scoring functions. II. The testing of an empirical scoring function for the prediction of ligand-receptor binding affinities and the use of Bayesian

regression to improve the quality of the model. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 503−519.
(60) 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.
(61) Korb, O.; Stützle, T.; Exner, T. E. Empirical Scoring Functions for Advanced Protein−Ligand Docking with PLANTS. *J. Chem. Inf. Model.* **2009**, *49*, 84−96.
(62) Berman, H. M.; Battistuz, T.; Bhat, T. N.; Bluhm, W. F.; Bourne, P. E.; Burkhardt, K.; Feng, Z.; Gilliland, G. L.; Iype, L.; Jain, S.; Fagan, P.; Marvin, J.; Padilla, D.; Ravichandran, V.; Schneider, B.; Thanki, N.; Weissig, H.; Westbrook, J. D.; Zardecki, C. The Protein Data Bank. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *58*, 899−907.
(63) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. Diverse, High-Quality Test Set for the Validation of Protein−Ligand Docking Performance. *J. Med. Chem.* **2007**, *50*, 726−741.
(64) Sadowski, J.; Gasteiger, J. From atoms and bonds to three-dimensional atomic coordinates: automatic model builders. *Chem. Rev.* **1993**, *93*, 2567−2581.
(65) 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.
(66) Good, A.; Oprea, T. 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.
(67) *Pipeline Pilot*, version 7.5.2; Accelrys: San Diego, CA, 2009.
(68) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(69) Korb, O.; McCabe, P.; Cole, J. The Ensemble Performance Index: An Improved Measure for Assessing Ensemble Pose Prediction Performance. *J. Chem. Inf. Model.* 2915−2919.
(70) McGovern, S. L.; Shoichet, B. K. Information Decay in Molecular Docking Screens against Holo, Apo, and Modeled Conformations of Enzymes. *J. Med. Chem.* **2003**, *46*, 2895−2907.
(71) Tuccinardi, T.; Botta, M.; Giordano, A.; Martinelli, A. Protein Kinases: Docking and Homology Modeling Reliability. *J. Chem. Inf. Model.* **2010**, *50*, 1432−1441.