J Comput Aided Mol Des (2012) 26:737–748
DOI 10.1007/s10822-012-9551-4

# Pose prediction and virtual screening performance of GOLD scoring functions in a standardized test

John W. Liebeschuetz · Jason C. Cole · Oliver Korb

Received: 28 September 2011 / Accepted: 9 February 2012 / Published online: 28 February 2012

**Abstract** The performance of all four GOLD scoring functions has been evaluated for pose prediction and virtual screening under the standardized conditions of the comparative docking and scoring experiment reported in this Edition. Excellent pose prediction and good virtual screening performance was demonstrated using unmodified protein models and default parameter settings. The best performing scoring function for both pose prediction and virtual screening was demonstrated to be the recently introduced scoring function ChemPLP. We conclude that existing docking programs already perform close to optimally in the cognate pose prediction experiments currently carried out and that more stringent pose prediction tests should be used in the future. These should employ cross-docking sets. Evaluation of virtual screening performance remains problematic and much remains to be done to improve the usefulness of publically available active and decoy sets for virtual screening. Finally we suggest that, for certain target/scoring function combinations, good enrichment may sometimes be a consequence of 2D property recognition rather than a modelling of the correct 3D interactions.

**Keywords** Docking · Enrichment · Pose-prediction · Virtual screening · GOLD · Scoring function

**Abbreviations**

| | |
|---|---|
| CCDC | Cambridge Crystallographic Data Centre |
| PDB | Protein Data Bank |
| RMSD | Root mean square deviation |
| VS | Virtual screening |
| ROC | Receiver operating characteristic |
| AUC | Area under curve |
| DUD | Directory of useful decoys |

## Introduction

The GOLD docking program has been in existence now for over 15 years [1, 2]. Despite its apparent seniority, continuous evolutionary development over the intervening years has kept it one of the most widely used and cited docking programs currently available ([3], an in house ISI Web of Science search on the primary articles showed GOLD to be the second most highly cited docking program in 2011). Where GOLD originally had only one scoring function, Goldscore, it now has four that can be used for docking (Goldscore [2], Chemscore [4, 5], ASP [6] and ChemPLP [7]). The challenge presented by the organisers of the 241st ACS meeting Docking and Scoring Symposium not only allowed the opportunity to test the performance of GOLD against standardized test-sets, under standardized conditions; but also to test the relative merits of the four scoring functions.

Each of the four scoring functions is significantly different in construction to the others.

- Goldscore has a Van der Waals treatment of clash and dispersion terms and uses a crystal structure derived treatment of hydrogen bonding and metal terms.
- Chemscore [4, 5] is an empirical scoring function parameterised from 82 complexes of known binding affinity; it has a lipophilic–lipophilic contact area term, a geometrically constrained hydrogen bond term and a term to penalise excessive flexibility.

J. W. Liebeschuetz (✉) · J. C. Cole · O. Korb
Cambridge Crystallographic Data Centre, 12 Union Rd,
Cambridge CB2 1EZ, UK
e-mail: john@ccdc.cam.ac.uk

- ASP is a knowledge-based scoring function derived from the protein data bank (PDB) [6].
- ChemPLP [7] is the most recently introduced scoring function. This treats neutral and repulsive contacts with a piecewise linear potential (PLP). This simple potential has both an attractive and repulsive part for neutral contacts and solely a repulsive part for anti-complementary contacts (donor–donor, metal-donor and acceptor–acceptor). The Chemscore hydrogen bonding term is used for hydrogen bonds and the Chemscore internal energy term is also used. ChemPLP is fast to calculate in comparison to the Gold score (benchmarked at 23 s for 10× repeat dockings at default settings, compared to Goldscore, benchmarked at 90 s).

## Experimental section

### Pose-prediction experiments

Protein and ligand structures were supplied for all 85 targets in the Astex Diverse Set [8]. In some 38 cases there are multiple sites of binding for the ligand. In such cases docking was carried out to all such binding sites even where such a site would not be considered the ‘normal’ binding site for the ligand. Twenty five protein structures were designated by the organisers to have either, (a) alternate conformations for the ligand or for side chains near the ligand (11), (b) crystal packing interactions with the ligand, (c) incomplete electron density for the ligand (4). Structures where these problems occurred are designated here the ‘rejected list’. The remaining 60 structures are designated here the ‘white list’.

Docking was carried out using GOLD 5.0.1 [9]. Binding sites were defined as the residues with at least one heavy atom within 6 Å (standard default) from the cognate ligand placement. No water was present in any binding site. The default docking protocol was applied (1.0× auto settings, 10 GA) and the best pose saved. Each experiment was then repeated 25 times. This protocol was repeated for the four scoring functions Goldscore, Chemscore, ASP and ChemPLP.

Success rates were calculated as the percentage of the entire dataset in which poses within the cut-off criterion with respect to the experimental pose were achieved. The primary cut-off criterion was 2 Å RMSD, although results using a more stringent 1 Å criterion are also presented.

Relibase+ [10], a tool for analysis of protein–ligand complexes, was used in cases where a detailed comparison of the experimentally determined binding modes was carried out.

### Virtual screening experiments

Binding sites were defined as the residues with at least one heavy atom within 6 Å (standard default) from the ligand in the supplied protein structure unless it was deemed that a larger binding site needed to be defined, either because the cognate ligand was small or the likely binding region was deemed to extend beyond 6 Å from the ligand. A binding site definition of 6 Å around the cognate ligand for all proteins was used in an additional experiment for one scoring function (ChemPLP). Water molecules in the active site, if deemed they could be involved in ligand binding, had hydrogen positions optimized during docking. For the three targets dhfr, gpb and pde5 it was deemed that some waters might be displaced on binding of certain ligands and, using the specific functionality for this [11] these were allowed to toggle on or off and, if ‘on’, were allowed to optimize hydrogen positions. They were also allowed to move up to 1 Å from starting positions, an option that became available in GOLD 5.0.

Fifty percent search efficiency auto settings were used for docking (i.e. 0.5X default). Each run was repeated ten times and the top pose in each case used in the enrichment calculations. This is a relatively slow protocol for virtual screening purposes so the experiment was repeated for one scoring function (ChemPLP) using 10% search efficiency.

Virtual screening experiments were carried out using all four GOLD scoring functions, Goldscore, Chemscore, Astex Statistical Potential (ASP) and ChemPLP. Docking was carried out on 40 DUD active/decoy sets [12] and 10 Wombat [13] active sets that appropriately corresponded to DUD active decoy sets, alongside the corresponding DUD decoys.

The total area under the receiver operating characteristic (ROC) curve was used as the primary measure of enrichment success. Early enrichment metrics were obtained by calculating AUCs for the ROC curve at 0.1, 1 and 2% false positive rate (FPR), and in addition, by calculating enrichment factors over the top 0.1, 1 and 2% of the database as ranked by score of the top-ranked pose in each case. Other metrics recorded were the median, the maximum and the minimum AUCs at 100% FPR.

AUCs and enrichment metrics were averaged over all the DUD sets to assess which scoring functions had best overall performance. In addition, the DUD sets were divided into the following constituent target classes, *nuclear hormone receptors* (8 members), *kinases* (9 members), *serine proteases* (3 members), *metalloproteases* (4 members), *folate enzymes* (2 members), and *others* (14 members) and the enrichment metrics calculated for these families separately.

Lastly, a Null experiment was carried out for which matched ligand and decoy sets were docked into a protein structure of a different target with similar binding site characteristics and size. This was carried out for all the DUD ligand-decoy sets according to pairings provided by the organizers. Often the paired proteins were of the same target class.

## Results and discussion

### Pose prediction

Table 1 shows the relative pose prediction performance of GOLD for the four scoring functions over the entire set of proteins. Success rates labelled 'Top Ranked' are calculated using only the best scored pose out of the 25 saved for each binding site. Success rates in the columns labelled 'Closest' measure in what percentage of cases at least one pose out of 25, satisfies the success criterion. This figure is useful when employed alongside the top ranked success rate because it is an estimate of the maximum achievable top ranked success rate assuming a perfect 'ranking' scoring function.

In all cases we present two sets of results; namely those that treat each binding site in each protein as a separate observation (All Sites) and those which give the best possible result we could have obtained by selecting one binding site for docking (Best Sites). The latter results are broadly comparable to the results obtained by Verdonk et al. [8]. In that study, the authors selected a binding site by visual inspection: if there were no obvious reasons for selecting one site over another, then the first site in the structure was used, but in cases where there were obvious reasons to choose one site over another an intelligent choice was made.

In all cases the top-ranked success rate is not substantially below the closest success rate at the 2.0 Å RMSD threshold (within 7–16% in all cases). The difference is larger for success rates at the far tighter 1.0 Å RMSD threshold (within 15–28% in all cases). ChemPLP achieves numerically the highest closest success rate at both the 2.0 and 1.0 Å cut-offs (91% over all sites at 2.0 Å, 76% at 1.0 Å).

ChemPLP numerically was the best performing scoring function according to top-ranked success rate (81% at 2 Å RMSD over all sites). Also notable are the results obtained at 1.0 Å RMSD. ChemPLP achieves a success rate of 59% at the threshold for all sites and 68% for the best site. This significance of this result was analysed using a binomial test to compare each pair of scoring functions. For each pair of scoring functions a set of entries was constructed that listed all binding sites that failed for one of the scores. For example, at the 2 Å RMSD threshold, for ChemPLP versus Goldscore, 36 entries were found; 26 entries succeeded with ChemPLP, the other 10 succeeded with Goldscore. A binomial test indicates that the probability of this occurring by chance is approximately 1% ($p$ value = 0.011) so we can have a high confidence that ChemPLP is out-performing Goldscore. For the other scores, the result is less clear-cut: ChemPLP out-performs ASP at the 95% confidence level with a $p$ value = 0.035. Chemscore is only significantly worse at a confidence level of 85% with a $p$ value = 0.14. These results suggest that ChemPLP is the most successful scoring function in GOLD for pose prediction. For comparisons at the 1.0 Å RMSD threshold, the $p$ values were ChemPLP:Goldscore = 0.14, ChemPLP:Chemscore = 0.013 and ChemPLP:ASP = 0.001. Consequently, further analysis in this work focuses on results achieved with ChemPLP.

Achieving poses within 1 Å RMSD rather than 2 Å would better allow correct ranking of that pose against those of other ligands, whether the ranking method uses the docking scoring function itself or another methodology. So, there is a strong argument that this more stringent criterion should be used in future. However this can only be done if experimental errors in placement of ligand atoms within the protein structures of the test set are sufficiently low.

Pose prediction performance for ChemPLP is shown in Table 2 for the 'white list' structures versus 'rejected list' structures and for 'all' and 'best' sites. Numerically, it appears that the 'rejected' list outperforms the 'white list'

**Table 1** Success rates for GOLD with the four scoring functions

| | All sites (154 sites) | | | | Best site only (85 sites) | | | |
|---|---|---|---|---|---|---|---|---|
| | Top ranked (%) | | Closest (%) | | Top ranked (%) | | Closest (%) | |
| | 2.0 Å | 1.0 Å | 2.0 Å | 1.0 Å | 2.0 Å | 1.0 Å | 2.0 Å | 1.0 Å |
| ChemPLP | 81 | 59 | 91 | 76 | 87 | 68 | 93 | 80 |
| Goldscore | 69 | 50 | 82 | 68 | 78 | 58 | 88 | 74 |
| Chemscore | 76 | 48 | 87 | 66 | 82 | 55 | 91 | 74 |
| ASP | 72 | 44 | 86 | 61 | 79 | 53 | 89 | 71 |

**Table 2** Comparison of pose prediction success with ChemPLP for rejected list and white list structures

| | All sites (154 sites) | | | | Best site (85 sites) | | | |
|---|---|---|---|---|---|---|---|---|
| | Top ranked (%, N) | | Closest (%, N) | | Top ranked (%, N) | | Closest (%, N) | |
| | 2.0 Å | 1.0 Å | 2.0 Å | 1.0 Å | 2.0 Å | 1.0 Å | 2.0 Å | 1.0 Å |
| White list (106 sites in total) | 79 | 66 | 92 | 77 | 85 | 69 | 92 | 78 |
| Rejected list (48 sites in total) | 84 | 42 | 86 | 72 | 92 | 64 | 96 | 84 |

**Table 3** Sampling analysis for each scoring function

| | % Of poses within 2.0 Å RMSD | | |
|---|---|---|---|
| | All structures | As % of closest success rate | Challenging structures |
| ChemPLP | 77 | 85 | 59 |
| Goldscore | 70 | 85 | 48 |
| Chemscore | 71 | 81 | 50 |
| ASP | 68 | 79 | 44 |

at the 2.0 Å threshold, but it is questionable whether this is significant within the bounds of uncertainty, as the rejected list contains just 25 structures (43 sites): The difference is just 5%, which equates to only 2 sites.

At the 1 Å cut-off the success rate for top ranked poses is better for the 'white list' structures. This is particularly true when all binding sites are considered. This is a more expected result: we would expect the difference between 'white list' and 'rejected' to become clearer at the lower cut-off. The difference is not very large for the subset of best sites but is considerable if we consider all sites. A possible explanation is that some of the additional binding sites in the 'rejected list' structures have poorly refined positions for the ligand. In support of this, the 20 sites that fall into the RMSD range of 1.0–2.0 Å are spread across only 11 structures, with code 1u1c supplying 6 of the sites. It is notable that this set features some sites where there is known disorder in the binding sites (1sg0, 1ig3 and 1tz8).

The large number of docking simulations performed (25 repeats per binding site and per scoring function) provides us with a window into the overall sampling performance of GOLD. For this purpose we define an additional subset of 'Challenging' structures. Challenging structures are defined as those where at least one scoring function fails to achieve a 100% success rate for pose retrieval.

The percentage of poses generated that were within 2.0 Å with each scoring function was recorded. These retrieval rates are given for the four scoring functions in Table 3. As is apparent, ChemPLP out-performs the other three scoring functions retrieving, on average, 76.6% of poses within 2.0 Å of the experimental position for the complete test set. For 'Challenging' structures a slightly larger performance difference is observed. If we examine only the cases in the full dataset where the experimental binding pose (within 2 Å RMSD) is reproduced at least once (c.f. the second column in Table 3), then ChemPLP shows superior sampling to Chemscore and ASP, and similar sampling to Goldscore

## Variability of binding sites within single crystal structures

A standard deviation for pose RMSD for each protein model was calculated over all binding sites for that model using the top-ranked poses obtained with ChemPLP. In nine cases standard deviations were observed that were greater than 0.5 Å. For protein models with more than one site, the average by-site standard deviation observed was 0.38 Å. The eight structures that have a high standard deviation are shown in Table 4. The results for some of these structures will be analysed in the next section.

Results are shown in Fig. 1 where the RMSD of the closest observed solution is plotted against the top ranked solution RMSD. We can separate docking simulations into three different categories, namely those that were successful, those that failed to attain a top-ranked solution within 2.0 Å but do generate a solution in the final ranked list within 2.0 Å (Ranking Failures) and those where no solutions are observed within 2.0 Å (General Failures).

**Table 4** Cases where there are significant differences between pose RMSD for different sites on a protein

| PDB code | Lowest top-ranked RMSD (Å) | Highest top-ranked RMSD (Å) | Top-ranked standard deviation (Å) |
|---|---|---|---|
| 2bm2 | 0.62 | 6.95 | 2.63 |
| 1u4d | 0.76 | 5.92 | 2.58 |
| 1ig3 | 0.39 | 5.27 | 2.44 |
| 1g9v | 1.14 | 4.59 | 1.72 |
| 1sq5 | 1.47 | 4.50 | 1.20 |
| 1q1g | 0.68 | 2.18 | 0.56 |
| 1tz8 | 0.77 | 1.97 | 0.56 |
| 1mmv | 0.58 | 1.58 | 0.50 |

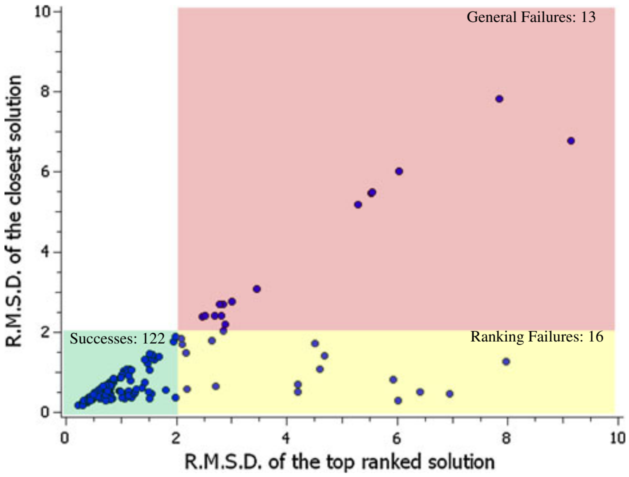

**Fig. 1** Distribution of RMSDs for top-ranked and closest solutions for solutions generated with ChemPLP. As is apparent, two kinds of test suite failure are observed. Those that can be deemed general failures, where no solutions are observed that are within 2.0 Å RMSD and ranking failures, where solutions that could be deemed successes are generated but not correctly ranked

Sixteen binding sites are found to be 'General Failures' when using ChemPLP. The 16 sites come from 10 distinct structures, namely 1hvy,1xm6,1oq5,1tz8,1gm8,1ig3,1jd0, 1q41,1uml and 1jje. Four of these, 1hvy, 1tz8,1ig3 and 1jje are listed on the 'rejected' list. 1oq5 and 1jd0 are two general failures which were latterly identified as being problematic for GOLD due to the protonation state of the input ligands provided in the test set. The protonation state is correct for the free ligand, but the ligand loses a proton on binding to a metal centre in the binding site and this is in not automatically accounted for with GOLD. Therefore we can unequivocally say these are failures of the docking program.

In 1tz8 there are three binding sites, one appears to be a true binding site. The other two are symmetry related sites between protein chains from different crystal unit cells and the ligand shows disorder in both sites. These are likely to be low affinity sites which are artefacts of crystallisation and arguably that they should not be in the test set. 1ig3 is another of the ten proteins for which disorder is present in the experimental structure.

The entry 1jje is worthy of note as an exemplar of a problematic case for any docking program. 1jje is on the 'rejected' list due to the presence of packing contacts to the bound ligand. The ligand binds via two carboxylate fragments that chelate metal centres in the protein. These interactions are well modelled by GOLD, but as is apparent from Fig. 2, the ligand is placed in a pose that is rotated in comparison to the observed binding mode. GOLD places the dioxoline ring incorrectly, but this is located in the region where the packing moieties are observed in the actual structure. Moreover, the ligand has pseudo-symmetry that allows key binding elements in experimental and docked poses to superimpose. Consequently the docked pose is an equally valid candidate for the binding pose seen in solution.

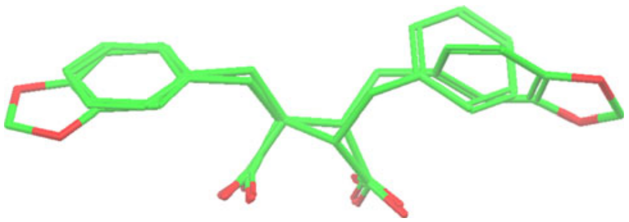

**Fig. 2** 1jje, overlay of the top-ranked GOLD solution and the observed binding mode using ChemPLP

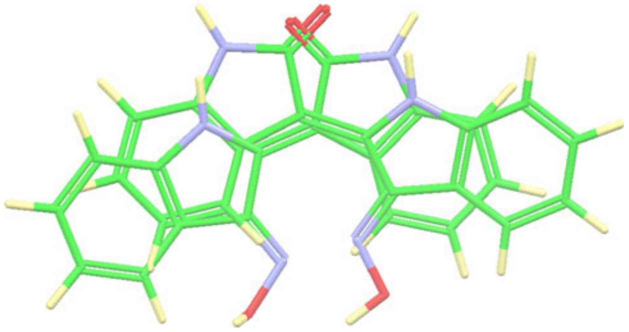

**Fig. 3** Docked and experimental binding modes for 1q41. These symmetry related binding modes are reasonable and are difficult to separate

Entry 1q41 is only listed as a general failure in one of its two binding sites, the second site is listed as a ranking failure. GOLD does dock the ligand in the correct location within the pocket, but rotated through 180°. The ligand is a kinase hinge binder and forms the classical interactions to the hinge motif. As such, it is entirely reasonable to expect binding with the ligand rotated due to the symmetric acceptor–donor–acceptor pattern of the motif. The overlay of the true binding mode and top ranked GOLD pose is shown in Fig. 3. The experimental structure forms an additional water-mediated contact from the ligand oxime to the protein via water 5602.

This result introduces the discussion of the role of water. No water is included in any of the binding sites and yet there are many cases where in the experimental structures, water mediated interactions between ligand and protein exist and may influence the precise binding mode found. This is an extremely difficult problem to deal with in a pose prediction evaluation because if most or all of the waters are left in, the template effect they provide is such that no docking program can fail to get the right result [8]. Nevertheless leaving water out completely may make it very difficult for any program to legitimately achieve success.

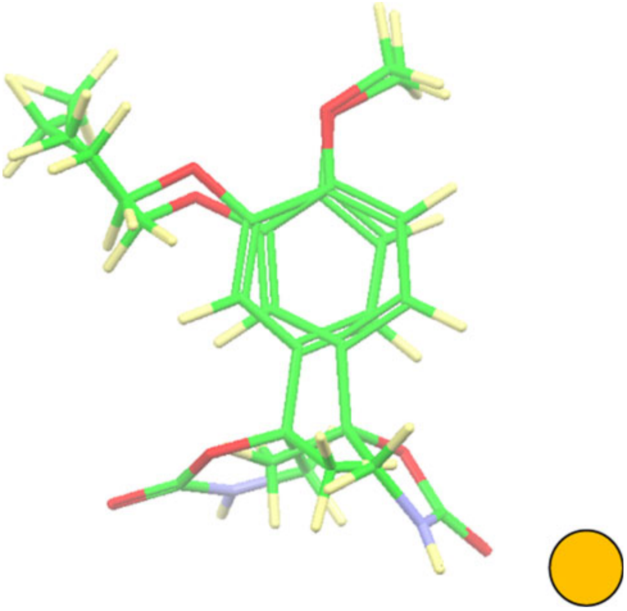

**Fig. 4** Docked and experimental binding modes in 1xm6. The *orange circle* represents the position of a water molecule which mediates an interaction with the ligand in the experimental structure

We can identify several cases other than 1q41 where docking failures and water mediated interactions are associated. Docking into 1xm6 results in a best pose with RMSD 1.2 Å. The difference between the docked pose and the correct pose is shown in Fig. 4. A water mediated interaction causes the imidazoline ring to rotate to the right in the experimental structure whereas it is found rotated left in the docked model. Binding of the ligand in 1gm8 is also dominated by three water mediated contacts between the ligand and waters 457, 458 and 459 in the binding site. Several other multiple water paths are also observed.

Some of the variability in docking between sites in Table 4 may also reflect the influence of water. In 1g9v, binding in both sites is mediated by two water molecules which are not included in the docking experiments. Consequently the experimental poses effectively have carboxylate groups that do not form strong interactions with the protein. In the predicted structures, the high RMSD solution has the carboxylate placed in the pocket of one of the water molecules. In 1u4d, for one site, 100% of docked poses are within 2 Å of the observed pose. In the other site, only 4% are within 2 Å, with the top ranked pose having a high RMSD. Only a score difference of 0.6 separates the top-ranked poses in each site. In the PDB structures water mediated contacts are observed for both sites but the water networks differ.

## Virtual screening

Figure 5 shows the histogram of AUC values at 100% FPR for all scoring functions over all DUD targets. It is clear there is considerable variation in enrichment success over the different targets with extremely good enrichment achieved for some targets and no enrichment (AUC $\leq 0.5$) for others. Although poor success for a given target may be because of a poor docking and scoring protocol, other possibilities, such as a poor choice of protein structure where protein mobility is likely; or incorrectly prepared active ligands, will also contribute to poor enrichment. Other docking programs that were tested using the same test set were found to show a similar distribution of poor and good enrichment and it would be interesting to see whether the problem cases were the same for all programs.

For many targets the performance is similar over all four scoring functions. However there are some cases where one scoring function performs better than the others. Table 5 shows the average enrichment metrics for all four scoring functions and early enrichment AUC metrics for different fractions of the false positive rate. Figure 6 gives the histogram for the same figures including 95% confidence limits on the means. Table 6 shows average enrichment factors for the top 0.1, 1 and 2% of the ranked datasets.

These results taken overall suggest that, for this dataset, ChemPLP has a slight edge over the other scoring functions for virtual screening, although ASP performance is not statistically significantly worse. The early enrichment metrics, 1 and 2% FPR AUC, show ChemPLP statistically outperforming both Chemscore and Goldscore.

Table 7 shows full statistical information for runs carried out using ChemPLP. Each run was repeated five times to assess variability in enrichment rates on repetition. Insignificant variation in enrichment statistics is found between runs.

Table 8 compares three different docking protocols for ChemPLP. The first is that described above. The second differs in using a much shorter virtual screening protocol employing 10% search efficiency settings. Despite the less efficient search protocol very similar enrichment metrics are found. The third protocol differs from the first in using a binding site including all residues within 6 Å (the normal default) of the cognate ligand of the protein model, rather than a binding site size whose distance criterion is defined by inspection. Where the binding site sizes were different, the site determined by the third protocol was always smaller. The enrichment metrics are slightly better when using the smaller binding site definitions. This is likely to be due to fewer high scoring poses being found for the decoys in some of the binding sites which are reduced in size. The alternative explanation, that the reduced size binding sites are searched more efficiently, can be discounted because we see that the 10% efficiency protocol works almost as well as the 50% efficiency protocol. Therefore sampling is not likely to be an issue in the third protocol.

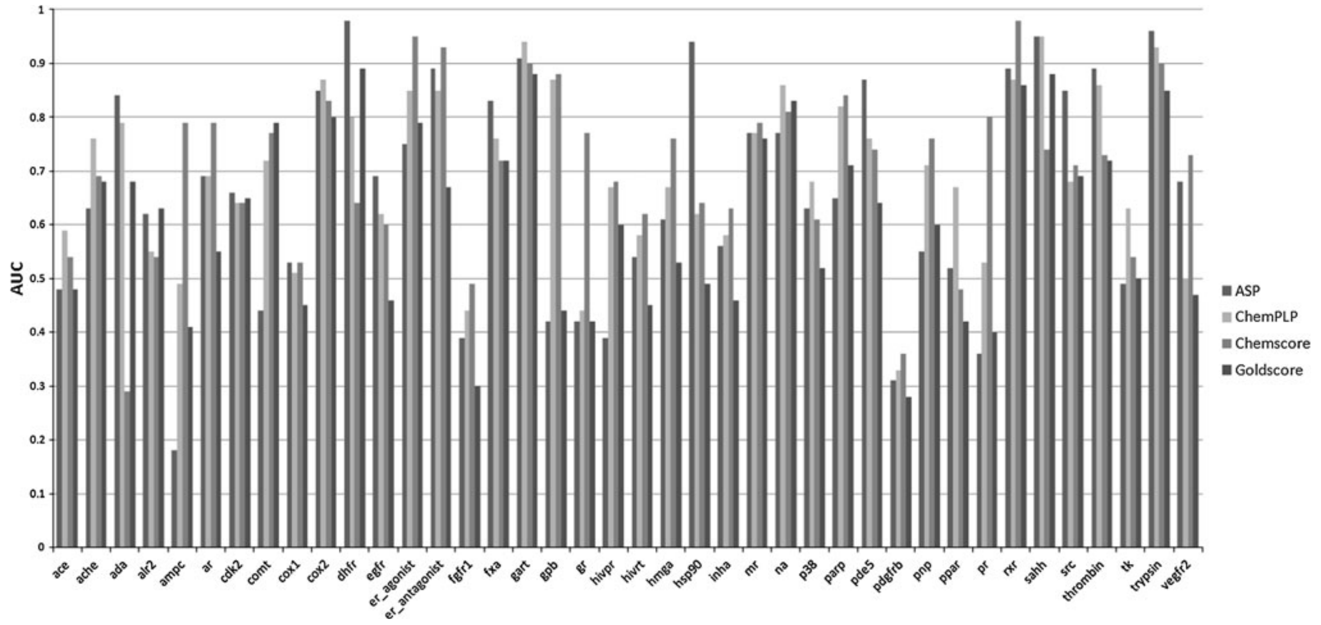

**Fig. 5** ROC AUCs for all 40 DUD active/decoy sets, over the four GOLD scoring functions

## Breaking down virtual screening results by protein class

Although ChemPLP appears the best scoring function overall it may be that other scoring functions are better for different target classes within the 40 protein targets. However it is essential to bear in mind the sample sizes may be small for some target classes. Figure 7a–e show the average early enrichment (AUCs at 2% FPR), and AUCs at 100% FPR, for five different subclasses of protein. Figure 7f gives the same figures for the remaining proteins in the DUD set. The error bars on the graphs represent the 95% confidence limits on the mean.

The small data sets preclude strong conclusions being made. However, taking both early enrichment and 100% AUCs into account these results do suggest that different scoring functions may be better suited to different enzyme classes. Chemscore for instance, works well on the whole for nuclear hormone receptors, whereas ASP and ChemPLP appear to be the functions of choice for serine proteases. ChemPLP is the scoring function of choice for those proteins not belonging to any of the above classes. Performance is worst overall for the kinase set of targets, perhaps reflecting that the significant protein mobility exhibited by this class may require several protein models to be evaluated instead of one.

Are then these differences in performance attributable to particular sensitivity of one scoring function to the general three dimensional arrangement of functionality within a class of receptors? To answer this question we turn to the results of the Null experiment.

## The Null experiment

The Null experiment tests the hypothesis that docking an active/decoy set into a binding site of a different target, but having similar size and polarity properties, leads to no enrichment. The assumption made is that the shape of the binding site will be sufficiently different between the members of a Null pair, so that an ideal docking protocol, using an ideal ligand/decoy set will generate Null 100% FPR AUCs that differ insignificantly from 0.50.

Table 9 compares the 100% FPR ROC AUCs between the Standard and the Null virtual screening experiments. The Null experiment provides AUCs that are on average, only slightly in excess of 50%, which is close to that expected. The difference in average AUCs between the Null and the Standard experiments can be taken to represent the discriminatory power of the docking programs at ranking actives according to the 3D shape and functional group disposition of the target protein, since in principle the paired proteins have similar sizes and ratios of polar to non-polar functionality. These differences are of the order of 0.13–0.16. ChemPLP is the most discriminatory scoring function by a small margin.

Although this appears a clean-cut result, further investigation is required because the average difference in AUCs only tell part of the story. It is clear from Table 9 that there is considerable variation in the difference of AUCs and there are cases where the difference is negative, i.e. retrieval is better against the null target! Figure 8 shows the histograms of Null AUC, normalised by subtracting 0.5.

**Table 5** Average AUC of the ROC curve for different fractions of false positive rate (An AUC at a FPR of 2% represents the portion of the ranked active-decoy list that contains the top 2% of the total number of decoys)

| % FPR | ROC AUC | | | |
|---|---|---|---|---|
| | 0.1 | 1 | 2 | 100 |
| ASP | 0.05 (±0.03) | 0.11 (±0.04) | 0.15 (±0.05) | 0.66 (±0.06) |
| ChemPLP | 0.08 (±0.03) | 0.14 (±0.03) | 0.17 (±0.04) | 0.70 (±0.05) |
| Chemscore | 0.03 (±0.02) | 0.08 (±0.03) | 0.12 (±0.03) | 0.70 (±0.05) |
| Goldscore | 0.06 (±0.03) | 0.11 (±0.03) | 0.14 (±0.03) | 0.61 (±0.05) |

Error ranges represent 95% confidence limits

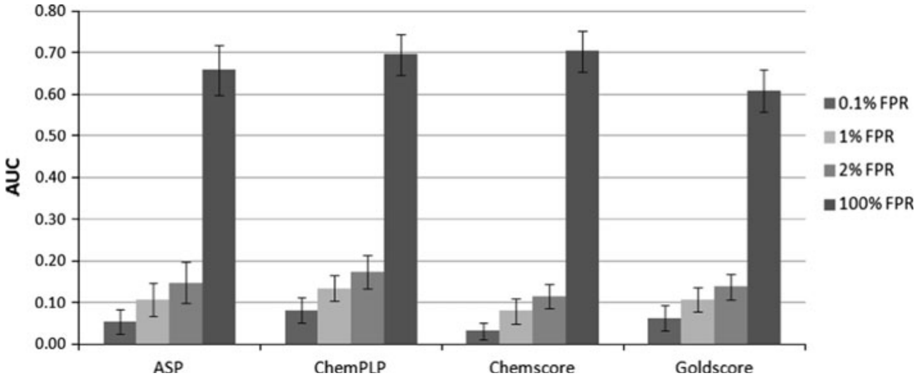

**Fig. 6** Average AUC of the ROC curve for different fractions of false positive rate. *Error* ranges represent 95% confidence limits

**Table 6** Enrichment factors calculated for the top 0.1, 1 and 2% of ranked active-decoy lists

| % Of dataset | Enrichment factor | | |
|---|---|---|---|
| | 0.1 | 1 | 2 |
| Max | 31 | 31 | 31 |
| ASP | 13 (±4) | 9 (±2) | 8 (±2) |
| ChemPLP | 20 (±4) | 13 (±2) | 9 (±2) |
| Chemscore | 10 (±4) | 9 (±2) | 7 (±2) |
| Goldscore | 17 (±4) | 10 (±2) | 8 (±2) |

Error ranges are 95% confidence limits

There are quite a number where the Null AUC is positive i.e. good retrieval is achieved against a ‘wrong’ target. There are also some cases where significant negative enrichment is achieved against the Null target. In theory, ideal decoys and active sets for 3D virtual screening should be perfectly matched to each other by having identical average size, functionality, and general shape distributions, only differing by the relative dispositions of functional groups on the surface of the active. An additional criterion is that no decoys would be potent actives for other drug targets related to the target protein. Such ‘perfect’ sets should always be retrieved at about 50% AUC in a well designed Null experiment. That we get substantially less than 50% AUC in some of the cases here, indicates that the actives and decoys are either poorly matched in terms of their size, functionality or shape, or, perhaps less likely, the decoy set is heavily populated with actives of the second target.

Some of the cases where good enrichment is obtained in the Null experiment will now be examined in detail (Fig. 8b,c,d). The serine proteases, factor Xa and thrombin were paired together (Fig. 8b). Reasonable enrichment is obtained for both proteins over all scoring functions in both the standard and the Null experiments (0.7–0.9 100% AUC). However this is a case of an unsuitable Null pairing. Serine proteases have fairly rigid active sites, and binding

**Table 7** ROC AUC enrichment metrics averaged over five runs for ChemPLP

| AUC (%) | Mean | Median | Max | Min |
|---|---|---|---|---|
| 0.1 | 0.08 (±0.01) | 0.07 (±0.002) | 0.37 (±0.015) | 0.00 |
| 1 | 0.14 (±0.001) | 0.12 (±0.005) | 0.49 (±0.01) | 0.00 |
| 2 | 0.18 (±0.001) | 0.17 (±0.000) | 0.57 (±0.005) | 0.00 |
| 100 | 0.70 (±0.001) | 0.69 (±0.0015) | 0.95 (±0.000) | 0.33 (±0.002) |

Error ranges are 95% confidence limits

functionality and sub-pocket shape are often conserved. Factor Xa and thrombin, both members of the blood coagulation cascade, are particularly close in structure, especially in the S1 pocket, and cross-activity at the micromolar level or lower is well known [14].

The other two pairs, estrogen receptor (agonists) and mineralocorticoid receptor (Fig. 8c); and epidermal growth factor (egfr) and heat shock protein 90 (Fig. 8d); also show anomalously high enrichments in the Null experiment but this is true for one scoring function only in each case. This scoring function is Chemscore for the nuclear hormone receptor pair, and ASP for the kinase pair. This result can be explained if we hypothesize that the contributions to the scoring function in these cases do not reflect specific position dependent interactions in the active site but instead 2D properties of the active molecules that are common to the actives of both proteins in the sets used here. If this is true then such 2D properties might be characteristic of other members of the same target class and so we would expect Chemscore to be a good scoring function in general for nuclear hormone receptors and ASP to be a good scoring function for other kinases. This is born out in Fig. 7a and b though in fairness ASP only shows strongly superior enrichment for the two kinases just mentioned. Chemscore however is clearly the most effective scoring function for other nuclear hormone receptors progesterone receptor, glucocorticoid receptor and androgen receptor.

Thus some enrichment seen in the Null experiment is likely to be due to the scoring functions carrying information that represents 2D characteristics of the ligands rather than specifically taking into account the quality of 3D interactions. This is not a surprising finding, scoring functions often contain elements that do not have a strong directionality dependency (e.g. the non-polar surface area contact term in Chemscore) and which could therefore stand in as 2D descriptors; and 2D methods have consistently shown themselves to be successful in virtual screening experiments, often more so than 3D methods [15]. It also does not invalidate the preferential use of these scoring functions to carry out a primary virtual screening campaign against a particular target. However it should be considered that the poses generated in these cases are less likely to reflect the true binding pose of the ligand and great care should therefore be taken if, as is often necessary, a further analysis or filtering is carried out on the basis of binding pose characteristics.

### Virtual screening against Wombat subsets

Virtual screening was also carried out for ten of the DUD protein targets using the Wombat ligand sets for these targets instead of the DUD ligands. These targets were alr2, ar, cdk2, cox2, egfr, fxaa, hivrt, p38, pde5 and ppar. The Wombat sets are in principle characterised by having a more even distribution; and in some cases a wider range, of chemotypes than the DUD active sets. The DUD decoy sets were used for these virtual screens.

Figure 9a and b compare the AUCs for the DUD and Wombat sets at 100% FPR and 2% FPR for all 10 targets. The 100% FPR AUCs, shown in Fig. 9a are significantly poorer for five targets, ar, cox2, p38, pde5 and ppar. The difference in AUCs for the ppar sets are especially striking. A better AUC is obtained for the Wombat alr2 active set. If we examine the early enrichment histogram (Fig. 9b) we find that performance is significantly worse for the Wombat sets ar, cox2 and pde5. However early enrichment against egfr and hivrt is improved.

The general conclusion is that, overall, the Wombat sets present a significantly tougher challenge than the DUD sets. Clearly the precise make-up of active and decoy sets can greatly affect the enrichment statistics.

## Conclusions

Designing good tests for comparing the relative performance of docking programs is still fraught with difficulty. The docking and scoring experiment described within this journal edition has valiantly attempted to remove many of the variables that prevent literature studies being comparable. Although on the whole we believe the attempt was successful, there was still room for obscuring variation between studies by different groups. In particular, one criterion left to the discretion of the participants was binding site definition. Although this is understandable as it may have been difficult to provide a definition that all

**Table 8** Mean ROC AUC enrichment metrics for three different protocols (a) Standard (50% search efficiency) averaged over 5 runs (b) 10% search efficiency (averaged over 5 runs) (c) standardized binding site size (6 Å around ligand) and 50% Search efficiency

| Protocol | Avg AUC 0.1% FPR | Avg AUC 1% FPR | Avg AUC 2% FPR | Avg AUC 100% FPR |
|---|---|---|---|---|
| 50% S/E. | 0.08 (±0.01) | 0.14 (±0.001) | 0.18 (±0.001) | 0.70 (±0.0005) |
| 10% S.E. | 0.07 (±0.002) | 0.13 (±0.0015) | 0.17 (±0.0015) | 0.70 (±0.0005) |
| 50% S.E. + default binding sites | 0.08 | 0.15 | 0.19 | 0.71 |

Error ranges are 95% confidence limits

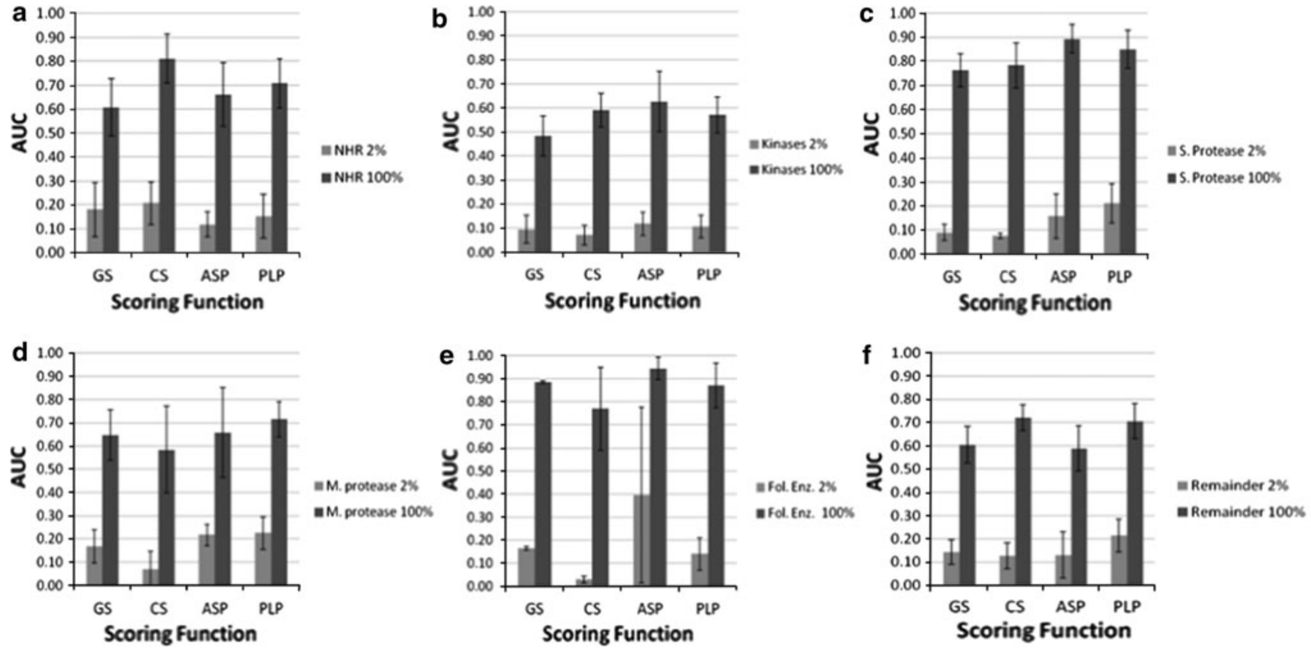

**Fig. 7** **a**–**f** Average early enrichment (at 2% FPR) and full ROC AUCs; **a** for nuclear hormone receptor subset of proteins ($n = 8$); **b** for the kinase subset of proteins ($n = 9$); **c** for the serine protease subset of proteins ($n = 3$), **d** for the metallo protease subset of proteins ($n = 4$); **e** for the subset of proteins where the inhibitors mimic folate ($n = 2$); **f** for the remaining proteins in the DUD set ($n = 14$). *Error* ranges represent 95% confidence limits

**Table 9** Difference in metrics between the Standard and the Null virtual screening experiment

| | Standard | Null | Standard minus Null | | Standard minus Null | | |
|---|---|---|---|---|---|---|---|
| | Avg AUC | Avg AUC | Avg AUC | Med. AUC | Lowest | Highest | SD |
| ASP | 0.66 (±0.06) | 0.52 (±0.07) | 0.14 | 0.15 | −0.34 | 0.64 | 0.22 |
| ChemPLP | 0.70 (±0.05) | 0.54 (±0.04) | 0.16 | 0.16 | −0.23 | 0.58 | 0.13 |
| Chemscore | 0.70 (±0.05) | 0.56 (±0.05) | 0.14 | 0.10 | −0.38 | 0.57 | 0.17 |
| Goldscore | 0.61 (±0.05) | 0.49 (±0.07) | 0.13 | 0.14 | −0.40 | 0.51 | 0.16 |

Error ranges are 95% confidence limits

docking programs were able to implement, binding site definition significantly affects both pose prediction and virtual screening success. In our study no binding site was defined to be smaller than 6 Å around the cognate ligand, the default value. Pose prediction experiments using a smaller binding site are liable to lead to artificially high success rates. On the other hand using too large a binding site definition may be detrimental. Where virtual screening was carried out on both a 6 Å binding site and a larger binding site we observed poorer enrichment statistics for the latter, on average.

Pose prediction success is necessary but not sufficient for good docking performance. This study has demonstrated that, of the scoring functions supplied with the GOLD docking program, the most recently introduced, ChemPLP [7], is the most effective scoring function for pose prediction in cognate protein–ligand complexes, achieving a success rate of 59% for all sites (68% for the best site) below 1.0 Å RMSD and 81% success over all sites (87% over best site per protein) below 2.0 Å RMSD. ChemPLP was also found the most effective scoring function in a recent study that looked at fragment docking [16].

High levels of accuracy in cognate pose prediction have therefore been achieved. It is not reasonable to expect a much higher level of success in this experiment, for two reasons. First, some of the sites were identified as being of questionable quality (the 'rejected list)' and three ligands were of the wrong stereochemistry. These errors could in principle be expunged by using a higher quality dataset. Second, but much harder to take account of, is the influence of water on ligand binding. No water molecules were

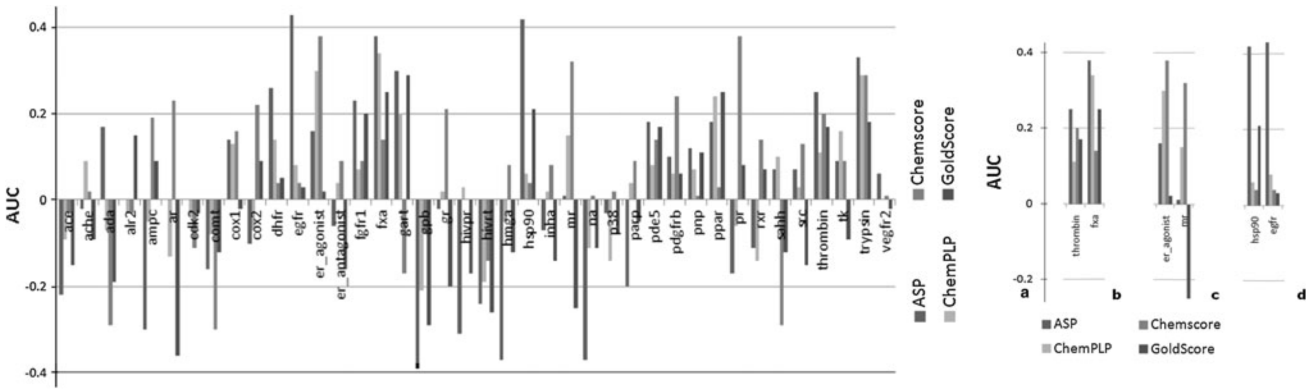

**Fig. 8** **a** Null enrichment ROC AUCs for all 40 DUD targets and four scoring functions, normalised so that the null value is 0; **b** Comparison of thrombin and factor Xa null AUCs, **c** Comparison of estrogen receptor (agonist) and mineralocorticoid receptor null AUCs, **d** Comparison of hsp90 and egfr tyrosine kinase null AUCs

included in the provided structures. Whilst this is a very reasonable pre-condition to impose as excessive water inclusion is liable to make pose prediction facile [8], we have found a number of cases in the test set where water mediated interactions play an important role in deciding the precise placement of the ligand. Consequently it cannot be expected that successful pose prediction will be achieved for these and other cases and therefore the maximum success rate achievable by any docking protocol in this experiment should be significantly less than 100%.

A different argument is whether the pose prediction test is actually relevant to standard praxis where docking is normally carried out against a non-cognate protein model. An experiment that involves cross-docking of ligands to other models of the target would appear to be a more realistic way of comparing the abilities of algorithms to generate approximately correct poses. It is even possible that the current practice of optimising performance of a docking algorithm for cognate pose prediction may lead to a degradation of performance of the same algorithm in a cross-docking experiment due to a greater sensitivity of such algorithms to small errors in protein atom placement.

Although some progress has been made [17, 18], few cross-docking test-sets that capture many different targets, and multiple ligands for those targets, yet exist. Success in published cross-docking experiments is generally shown to be much lower than in cognate docking [17, 18] and it is likely that new methodologies, perhaps involving induced fit around the ligand [19] or the use of multiple protein conformations [20]; will be needed to achieve satisfactory pose prediction success.

Measuring success at virtual screening is, if anything, a more difficult problem. A recurring issue is the construction of adequate and comparable active and decoy sets. A number of the DUD sets show less than 50% AUC in a Null experiment in this study and for 12 targets the Null AUC is less than 30% for at least one scoring protocol. This is a sign that the actives and decoys are not perfectly matched for some characteristics. Even where molecular weight distributions and functionality are similar for both actives and decoys, issues exist. For instance if the actives are largely of one chemotype then this could score particularly badly in the Null binding pocket, perhaps because of being forced out of a good conformation. The decoys, on the other hand might display a broad range of scores.

For five out of nine datasets where both DUD and Wombat subsets of actives are available, the ROC AUC in the standard docking experiment drops for the Wombat sets. The Wombat sets were designed to contain a more even distribution of chemotypes than the DUD sets and certainly appear to present a tougher virtual screening challenge. The experiments carried out here have neglected protein movement. Where a greater number of chemotypes can influence the overall ROC AUC it is possible, indeed may be expected, that good enrichment cannot be obtained if using a single protein model.

We should keep in mind that a docking program should succeed at virtual screening on the basis of correctly identifying and scoring the important 3D interactions that a ligand makes. However enrichments can also be achieved if scoring functions correlate with 2D properties of the ligand, which themselves, are characteristic of known inhibitors for that class of target proteins. The Null experiment allowed us to identify where this may occur. A key diagnostic is where only one scoring function out of the four tested shows enrichment in the Null test (if all four show enrichment it is more likely that the null pairing has put together two targets similar enough that cross-activity can occur). For instance Chemscore is a good scoring function for nuclear hormone receptors but we also find that it is also the only scoring function that shows good

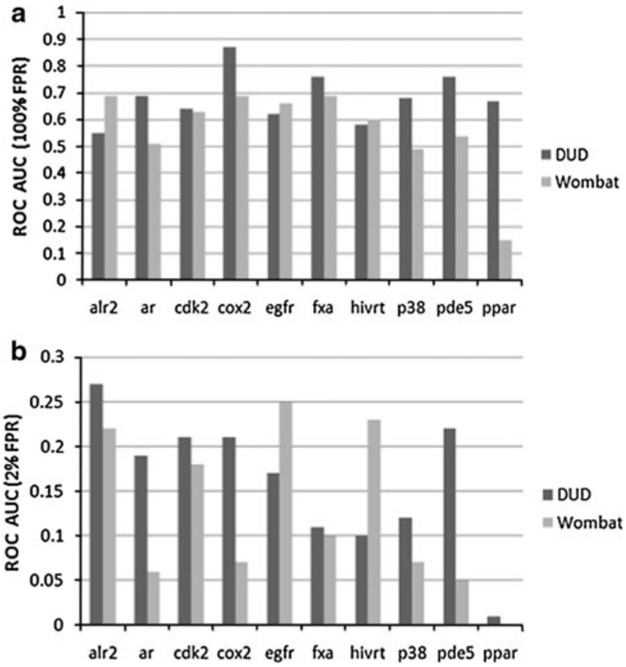

**Fig. 9** Comparison of performance on 10 comparable DUD and Wombat datasets using ChemPLP **a** AUC at 100% FPR, **b** AUC at 2% FPR

enrichment in the null experiment for the nuclear hormone pair, estrogen receptor and mineralcorticoid receptor.

The difference between the mean standard AUC and the mean null AUC may be taken to be indicative of the amount of 2D property recognition implicit in the scoring function. The lower the value, the greater is the 2D component of the recognition. In this regard, of the four scoring functions examined, ChemPLP is the one that suffers least from 2D molecular property bias.

Despite some of the already mentioned shortcomings of this study, we are able to establish that the most recently introduced scoring function in GOLD, ChemPLP, is highly effective for both pose prediction and virtual screening, and compares well in comparison with the other GOLD scoring functions Chemscore, ASP and Goldscore, outperforming them in a number of areas. It is also the scoring function appears to suffer least from 2D property bias. ChemPLP is fast to calculate (4× faster than Goldscore) and therefore suitable for virtual screening applications. ChemPLP would therefore appear to be the scoring function of choice for use within GOLD unless prior target-specific docking data suggests otherwise.

**Acknowledgments** 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. We also thank Dr Colin Groom for valuable comments regarding the manuscript.

## References

1. Jones G, Willett GP, Glen RC (1995) J Mol Biol 245:43–53
2. Jones G, Willett P, Glen RC, Leach AR, Taylor RJ (1997) Mol Biol 267:727–748
3. Sousa SJ, Alexandrino PS, Ramos MJ (2006) PROTEINS 65:15–26
4. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP (1997) J Comput Aided Mol Des 11:425–445
5. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD (2003) PROTEINS 52:609–623
6. Mooij WTM, Verdonk M (2005) PROTEINS 61:272–287
7. Korb O, Stützle T, Exner TE (2009) J Chem Inf Model 49(1):84–96
8. Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WTM, Mortensen PN, Murray CW (2007) J Med Chem 50:726–741
9. http://www.ccdc.cam.ac.uk/products/life_sciences/gold/
10. Hendlich M, Bergner A, Günther J, Klebe G (2003) J Mol Biol 326:607–620
11. Verdonk ML, Chessari G, Cole JC, Hartshorn MJ, Murray CW, Nissink JWM, Taylor RD, Taylor R (2005) J Med Chem 48:6504–6515
12. Huang N, Shoichet BK, Irwin JJ (2006) J Med Chem 49:6789–6801
13. Good AC, Oprea TI (2008) J Comput Aided Mol Des 22:169–178
14. Dönneke D, Schweintz A, Stürzebecher A, Steinmetzer P, Schuster M, Stürzebecher U, Nicklisch S, Stürzebecher J, Steinmetzer T (2007) Bioorg & Med Chem Lett 17:3322–3329
15. Bender A, Glen RC (2005) J Chem Inf Model 45(5):1369–1375
16. Verdonk M, Giangreco I, Hall R, Korb O, Mortensen P, Murray CW (2011) J Med Chem 54:5422–5431
17. Sutherland SJ, Nandigam RK, Erikson JA, Vieth M (2007) J Chem Inf Model 47:2293–2302
18. Verdonk ML, Mortenson PN, Hall RJ, Hartshorn MJ, Murray CW (2008) J Chem Inf Model 48:2214–2225
19. Sherman W, Day T, Jacobson MP, Friesner RA, Farid R (2006) J Med Chem 49(20):534–553
20. Jain AN (2009) J Comput Aided Mol Des 23:355–374