



[Artificial Receptors](#)

[Flexible Sidechains](#)

[GOLD Validation:
Virtual Screening](#)

[Optimum Docking &
Rescoring Protocols](#)

[Purchasing Information](#)

GOLD performance in the SAMPL Blind Virtual Screening Test

Introduction

In late 2007, OpenEye launched the SAMPL blind test of protein and ligand modelling. This set out to provide a truly objective test-bed of computational drug design methodology. A substantial part of the challenge was to test performance in Virtual Screening with few restrictions on the methodologies that could be used. Two protein targets were selected as test cases, JNK Kinase and urokinase. Structural models were available for both targets making both suitable for application of structure-based methods. The latter target is a serine protease and [previous work](#) at CCDC had already been undertaken to uncover a current best High Throughput Virtual Screening (HTVS) protocol for serine proteases. Therefore the SAMPL challenge appeared to be an appropriate forum in which to test the best candidate protocols.

Methodology

Docking, rescoring and post-processing was carried out using GOLD 3.2 and GoldMine1.1. The active site was modelled from the *1owd* structure supplied with the SAMPL ligand files. The protein was protonated using an in house implementation available now with GOLD 4.x. Non-conserved waters were removed and the binding site defined for GOLD as a region of 10 Å extent around the ligand in *1owd*. All ligands were docked using the GoldScore scoring function. Automatic Genetic Algorithm settings were used at 30% efficiency. Ten docking runs were carried out per ligand and the top three poses were saved. Use of the Diverse Solutions option was activated to ensure that every solution for a ligand was at least 1 Å RMSD from any other. No target specific restraints were applied. The ligand set comprised 8351 ligands. An additional rescoring protocol was also applied. The scoring function used for this was the Astex Statistical Potential (ASP). Poses were Simplex minimised with respect to the ASP scoring function, a default option for rescoring with ASP.

Ranked sets of ligands from three different post-processing protocols were submitted back to the SAMPL moderators. The first set comprised the ranking according to the GoldScore fitness term for the best pose per ligand. The second set comprised the ranking according to the rescore ASP fitness term for the corresponding best pose per ligand. This represents the [protocol](#) that was previously found most effective for HTVS against another serine protease factor Xa. The third set was ranked according to a consensus scoring function constructed from both the ASP and the GoldScore fitness terms for each pose. The fitness function terms were first normalised and then summed to generate the Consensus function. These operations were carried out using the Descriptor Arithmetic functionality available within GoldMine1.1.

Results

The active set turned out to comprise 27 actives in number. These were further subdivided into two sets, a set of nine "lead like" actives with low molecular weight and weak affinity; and a set of eighteen actives with good affinity. Results were returned showing the recovery of the "lead like actives" and "all" actives separately.

The decoy sets comprised several well known sets such as the Rognan set (990), the Schrödinger set (1000) and a set of DUD-like decoys. In addition a set of known inactives was included, and a set of active structures modified by addition of aromatic groups at positions designed to harm activity (labelled "clash" in tables). Finally two other decoy sets were created, firstly a set of structures (40 per active) with similar shape according their ROCS shape score (labelled "shape"), and a set of decoys made up of all other decoy structures, where each heteroatom that normally makes important electrostatic interactions, was replaced with carbon (labelled "uncolor"). This latter set contained some extremely non-druglike molecules of high symmetry, that GOLD had particular problems in dealing with. Where this was the case the structure was dropped from the list and placed last in the rankings. The decomposition of the decoy sets into different "flavours" allowed some in-depth analysis of virtual screening performance against different decoy types.

Three enrichment metrics were reported back to participants. Enrichment factors were calculated from a 2% top-cut, and a 10% top-cut of the ranked datasets. In addition areas under the Receiver Operating Characteristic Curve were calculated for the first 10% of the ranked list only. This then gives a measure of early enrichment. The maximum AUC under 10% ROC is 1.0 (100%) and the minimum is 0.0 (0%). Random enrichment should give an AUC under 10% ROC of 0.05 (5%). Tables 1, 2 and 3 report these enrichment metrics for the three different HTVS protocols submitted. The enrichments are broken down according to the "all" and lead-like ("lead") active sets. The data are broken down into performance against different decoy sets. The colour coding in the tables relates to the ranking of the protocol as compared to other protocols in the SAMPL challenge. "Red" indicates poor, 1st quartile ranking, orange indicates 2nd quartile ranking, blue, third quartile and green, 4th quartile ranking. Where it was not possible to deduce comparative performance from the data supplied by the SAMPL moderators, the text is black.

name	ROC auc(10%)	enrichment@2%	enrichment @10%
all-rognan	0.40	16.67	5.26
all-schrod	0.31	7.69	4.35
all-shape	0.45,	20.37	5.48
all-uncolor	0.50	20.97	5.93
all-clash	0.11	6.52	1.43
all-dud	0.32	8.70	5.00
lead-rognan	0.32	11.11	4.44
lead-schrod	0.22	5.00	3.33
lead-shape	0.36	16.67	4.44
lead-uncolor	0.38	16.67	4.29
lead-clash	0.03	4.17	0.91
lead-dud	0.24	5.56	4.29

{[uncolor](#)': 3937, '[lead](#)': 9, '[clash](#)': 18, '[dud](#)': 1895, '[schrod](#)': 1000, '[shape](#)': 723, '[inactive](#)': 7, '[rognan](#)': 990, '[active](#)': 18}

Graph 1. Enrichment metrics for GoldScore fitness alone.

name	ROC auc(10%)	enrichment@2%	enrichment @10%
all-rognan	0.71	29.69	8.57
all-schrod	0.55	25.00	6.47
all-shape	0.59	25.96	6.92
all-uncolor	0.81	38.10	8.57
all-clash	0.45	24.07	5.00
all-dud	0.48	19.87	6.67
lead-rognan	0.63	22.73	8.00
lead-schrod	0.41	16.67	5.56
lead-shape	0.46	16.67	6.25
lead-uncolor	0.75	37.50	8.00
lead-clash	0.29	16.67	3.33
lead-dud	0.34	11.54	5.71

{[uncolor](#)': 3937, '[lead](#)': 9, '[clash](#)': 18, '[dud](#)': 1895, '[schrod](#)': 1000, '[shape](#)': 723, '[inactive](#)': 7, '[rognan](#)': 990, '[active](#)': 18}

Graph 2. Enrichment metrics for GoldScore poses rescored with ASP.

name	ROC auc(10%)	enrichment@2%	enrichment @10%
all-rognan	0.55	25.00	6.32
all-schrod	0.49	22.22	5.56
all-shape	0.54	25.00	5.93
all-uncolor	0.64	29.41	7.83
all-clash	0.36	19.85	4.07
all-dud	0.45	19.23	5.86
lead-rognan	0.47	20.00	5.56
lead-schrod	0.39	16.67	4.44
lead-shape	0.46	20.00	5.56
lead-uncolor	0.55	27.78	6.67
lead-clash	0.21	13.64	2.86
lead-dud	0.36	12.50	5.71

{[uncolor](#)': 3937, '[lead](#)': 9, '[clash](#)': 18, '[dud](#)': 1895, '[schrod](#)': 1000, '[shape](#)': 723, '[inactive](#)': 7, '[rognan](#)': 990, '[active](#)': 18}

Graph 3. Enrichment metrics for GoldScore + ASP consensus score.

Taking the results as a whole the GoldScore docking/ASP rescoring and ranking protocol appears overall to perform the strongest whereas the GoldScore docking/GoldScore ranking protocol performs the poorest. The consensus protocol has a performance that is intermediate between the other two. Significant differences are observed between the

different decoy/active combinations. In particular performance of all protocols is poorer at retrieving the lead-like set of actives than the full set (lead-like+drug-like) of actives, as might be expected. Performance in comparison with other methodologies is the same however and, for the best protocol, "lead" retrieval is good.

The results indicate very clearly that although docking and scoring solely with GoldScore is a protocol with worse than average performance compared with other methodologies, docking using GoldScore and then rescoring with ASP, is a protocol which performs well in comparison with many other methodologies. Until further details are published on the SAMPL study a closer comparison with other good performant protocols/methodologies is not possible. However it should be noted that a feature of the docking/rescoring protocol used here is that no significant biases or restraints were added to the protocol to take account of prior knowledge regarding favoured modes of interaction of the ligand with the protein.

It is striking that, although ranking a Virtual Screen according docking scores may lead to mediocre enrichment, ranking these same poses with a second scoring function can lead to greatly improved performance. This is particularly evident in the "all-clash" and "lead-clash" pairs of results, where GoldScore docking/ranking is only lowest quartile performant, whereas GoldScore docking/ASP ranking is 3rd/4th quartile performant. The most likely explanation is that reasonable binding poses of actives are being generated by GoldScore docking, but these are not being ranked correctly against inactive structures. The rescoring step leads to a much improved ranking. It is worth pointing out as a matter of interest that the [previous study](#) on factor Xa had demonstrated that docking and ranking with ASP lead to significantly poorer results than docking and ranking with GoldScore. The same study also demonstrated the importance of Simplex minimisation for good performance of the rescoring protocol.

Conclusions

The SAMPL study has provided a useful and welcome testbed for the evaluation of structure-based methodologies. A test of three different protocols in this study, has demonstrated that best results are obtained with a protocol that employs rescoring and ranking of GoldScore poses with the Astex Statistical Potential, using Simplex minimisation on rescoring. A consensus scoring approach also gave good results. However ranking using the GoldScore fitness figures only, gave poorer enrichment. The rescoring protocol gives similar performance to the other best performing structure-based methodologies FRED, ROCS and Glide, according to [press reports](#). The promising results obtained here, set alongside previous in-house work, suggest that rescoring with a second scoring function may have generality as a method of improving enrichments from GOLD virtual screening experiments. It is planned now to investigate a variety of other target types to establish appropriate high performing docking/rescoring combinations.

Jan 2009