

SimBioSys®

Technical Note

## Case Study: Docking validation using the Astex Diverse set

### Reference

Hartshorn, M.J.; Verdonk, M.L.; Chessari, G; Brewerton, S.C.W.; Mooij, 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.

### Study Overview

Both software developers and users of docking tools need high quality sets of receptor-ligand complexes for benchmarking, testing and comparing docking methods and scoring functions. Validation, or self-docking, studies are probably far from real world drug discovery or lead optimization scenarios, but they offer a very accurate and simple metric for performance evaluation – the RMSD of the docked poses from the crystallographic conformation.

In many comparative studies emphasis is given to the selection of targets and ligands so that the set will be as diverse as possible in terms of covered protein families, dominant interaction types, size and properties of ligands etc. Often, some filtering work is done to reassure that the crystal structures that are being used are relatively free of errors, and are accurate enough to guarantee meaningful comparative metrics. It is rare to find as detailed an examination of the complexes as that found in the paper from the Astex group listed above.

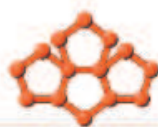
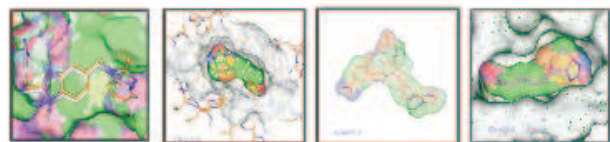
The constructed 85 complexes set is a subset of the original Astex/CCDC set of 305 complexes, which has been narrowed down by the following criteria:

- Targets can be represented only once.
- All receptors are of relevance to pharmaceutical or agrochemical purposes.
- Only structures for which the electron density map supports the entire ligand binding mode are considered.
- In all structures the ligand is in contact with protein atoms from only one copy of the biological system in the asymmetric unit.

In addition to the filtering, the authors also thoughtfully assigned hydrogen atoms to the ligands and receptors to provide a complete data set for validation studies.

### Methods

To accurately model a receptor-ligand system, one needs to identify the protonation states of the two entities which are determined by the interaction between the two. eHiTS takes a unique approach to the protonation problem by systematically evaluating all possible protonation states for both the receptor and ligand efficiently in a single run. The use of Interaction Surface Points (ISP) allows the assignment of ambiguous property flags for positions that could be either protonated or deprotonated. Then, during the docking run both states of such surface points are evaluated and scored, selecting the best protonation state for each individual interaction independently. This approach avoids the introduction of biases by pre-



SimBioSys®

## Technical Note

assigning hydrogen atoms, or the necessity to run several docking runs with different initial protonation states.

In addition, eHiTS 2009 allows running docking jobs with prefixed protonation states for the ligand, the receptor or for both. This is done using the “fixproto” flag. In this study, however, we focused on eHiTS’ mechanism of protonation handling to emphasize its strength as a rigorous and labor saving feature. We used the Astex Diverse set with eHiTS 2009, as well as with eHiTS 6.2. In both cases eHiTS was run with the default accuracy with the files provided by Astex and with no additional preparation.

The command line for both eHiTS 2009 and eHiTS 6.2 was:

```
ehits.sh -receptor receptor_file.mol2 -ligand ligand_file.mol2
```

To use the ligand and receptor protonation states provided by the Astex group one could use the following command line in eHiTS 2009:

```
ehits.sh -receptor receptor_file.mol2 -ligand ligand_file.mol2 -fixproto both
```

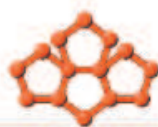
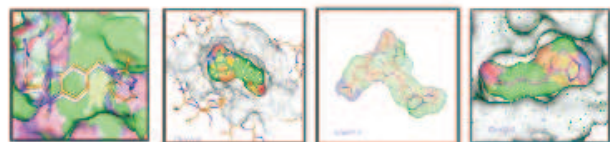
## Results

The success rates, given as the percentage of docked ligands for which the top-rank pose was within 2 Å of the crystallographic conformation, are given in the table below:

Software	Success rate (%)
eHiTS 9.0 with on-the-fly protonation handling	82
eHiTS 6.2 with on-the-fly protonation handling	72
GOLD – Standard protocol	81

The average run-time was 235 sec/ligand on an Intel Xeon™ Pentium-4 2.40GHz CPU, and 27 seconds on a Cell/B.E. processor. It should be noted that the dramatic 10 fold speedup on the cell processor is even exceeded in screening scenarios where the receptor is processed only once and the ligands are all docked to the prepared binding pocket.

The following charts demonstrate the distribution of the top-rank and the best (closest to crystallographic mode) poses according to RMSDs for the new eHiTS 2009 vs. eHiTS 6.2. While the quality of best poses has been negligibly reduced for this set in eHiTS 2009, clearly ranking of poses has improved with over 50% of the top-ranked poses being under 1 Å RMSD. Having better top-ranking structures is, of course, a more desirable feature in virtual screening and lead optimization scenarios where in the absence of adequate structural information only the top, or several top ranking solutions can be considered for further analysis.



SimBioSys®

## Technical Note

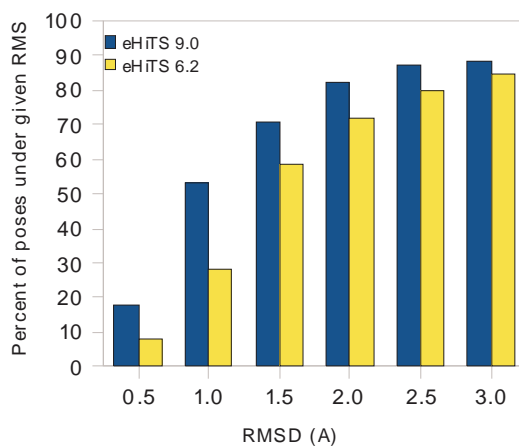


Figure 1 Top-rank poses distribution according to RMSD

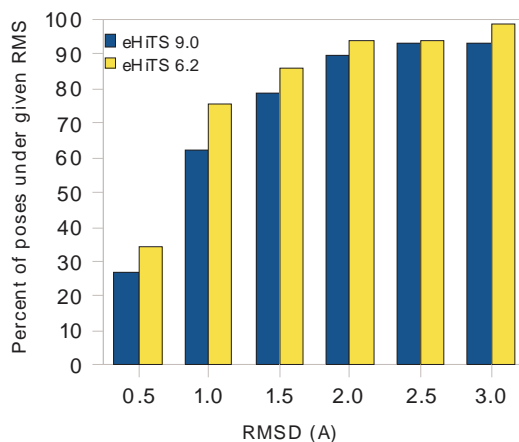


Figure 2 Closest poses distribution according to RMSD

This test case demonstrates along with others the strengths of eHiTS in pose prediction and ranking. This high level of accuracy can also be achieved at very high speeds using the Cell processor technology.

### Data

The Astex Diverse set has generously been made available to the public by the authors. It is available for download at: [http://www.ccdc.cam.ac.uk/products/life\\_sciences/gold/validation/astex\\_diverse/](http://www.ccdc.cam.ac.uk/products/life_sciences/gold/validation/astex_diverse/)