

## Case Study: Predicting Binding Affinities

### Reference

- [1] Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind Database: Methodologies and updates, *J. Med. Chem.*, **2005**, 48(12), 4111-4119.
- [2] Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures, *J. Med. Chem.*, **2004**, 47(12), 2977-2980.
- [3] Wang, R.; Lu, Y.; Fang, X.; Wang S. An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein-ligand complexes, *J. Chem. Inf. Comput. Sci.*, **2004**, 44, 2114-2125.
- [4] Raub, S.; Steffen, A.; Kamper, A.; Marian, C.M. AIScore - Chemically diverse empirical scoring function employing quantum chemical binding energies of hydrogen-bonded complexes, *J. Chem. Inf. Model.*, **2008**, 48, 1492-1510.

### Study Overview

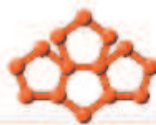
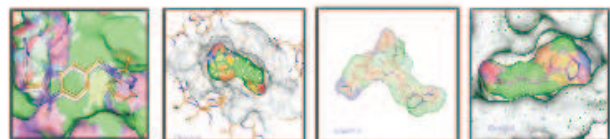
In order for molecular docking to become a driving force in drug discovery, it should demonstrate a predictive capacity with respect to binding affinity of the ligands to the receptor. Whether in screening large libraries of compounds, or in evaluating the effects of structural modifications to a lead molecule, a docking program is expected to provide valid top-rank poses for the species of interest and to rank them in proportion to their binding affinity, or at least to order them properly and reliably. The correlation between scores of top-rank poses of ligands to their experimental  $\log(K_d)$  values is therefore an important test for the outcome of the convolution between the pose prediction algorithm and the scoring function.

The PDBBind database [1,2], and primarily the “refined set”, provide an important tool for the evaluation, comparison and training of docking programs. It offers a wide selection of protein-ligand complexes from the PDB that have reliable measured values of the binding affinity, as  $K_d$ ,  $K_i$  and  $IC_{50}$ . The refined set consists of complexes that were solved at a resolution better than 2.5 Å, pertain to pharmaceutically relevant small molecules bound to binding pockets composed of naturally occurring amino acids, and for which  $K_i$  and  $K_d$  values were published. The PDBBind database and the refined set are updated annually to reflect new published data and enhanced refining protocols of the dataset.

We used the last versions of the refined set from 2008 to test the correlation of eHiTS 2009 scores with binding affinities in a virtual screening scenario, i.e. when no structural information is available with respect to the binding mode. This is different from other studies that have calculated the correlation only for successfully docked ligands with  $RMSD < 2.0$  Å.

### Methods

The entries in the refined set of the PDBBind database are available in pre-split form. The receptors are available in .pdb format, and the ligands are in .mol2 format. No additional preparation of the receptor or ligand is required for eHiTS, making this software exceptionally easy to use compared to other commercial packages. For



SimBioSys®

## Technical Note

each receptor/ligand pair we ran a docking job with eHiTS 2009 using the following command line:

```
ehits.sh -receptor protein.pdb -ligand ligand.mol2 -clip ligand.mol2 -rms ligand.mol2 -bindener
```

The **-receptor** and **-ligand** flags indicate the files for the respective molecules (when the pdb contains a complex, one may use the **-complex** flag, and then eHiTS would identify the ligand automatically and clip the receptor around it). The **-clip** flag specifies a file with coordinates that are used to determine the binding pocket.

By default eHiTS clips a box with a 10 Å margin from the receptor around the provided coordinates. The margin can be adjusted by the user using the **-margin** flag. In this case we clip the receptor around the ligand itself and use the default size for the clip box. eHiTS ran with the default accuracy mode (accuracy 3).

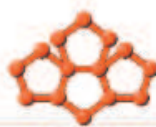
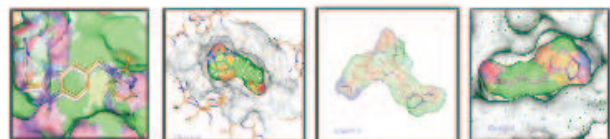
The **-bindener** flag should be used whenever the goal is to rank order active compounds. This option invokes a slightly modified scoring scheme that produces improved score – binding affinity correlations. The **-rms** is actually not necessary for the binding affinity calculation, as RMSDs of the poses are irrelevant for this task, but we included it to collect data about correlations between pose accuracy and binding affinity prediction, as shown below. The flag defines a ligand file to which the docked poses are compared. If binding affinity correlation were the sole purpose of this study, the following command line would have been sufficient:

```
ehits.sh -receptor protein.pdb -ligand ligand.mol2 -clip ligand.mol2 -toprank 1 -bindener
```

Where the flag **-toprank** indicates the number of best solutions to be saved as output. In this case the top rank solution is sufficient.

## Results

Out of 1401 cases, 31 (2.2%) did not dock in the binding pocket. Figure 1 shows the correlation between the binding affinity, given in pKd units, and the eHiTS score for the successfully docked ligands. The sign of the eHiTS score has been changed so that higher scores would reflect higher affinities, to conform with the convention adopted by previous studies [3,4].



SimBioSys®

## Technical Note

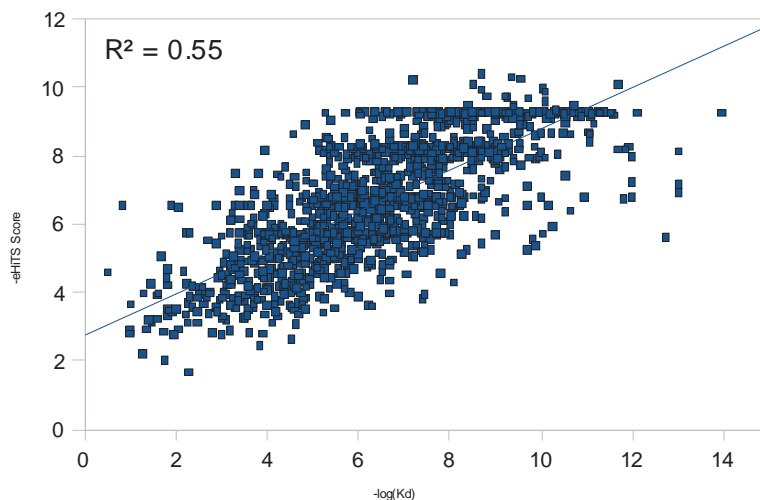


Figure 1. Correlation between eHiTS score and experimental binding affinity for the refined set of PDBBind-2008.

Overall, in spite of the scattering of the points, a clear trend is visible. The statistics of the linear regression model are given in the table below.

Scoring function	$R_p$	SD	ME
eHiTS-2009	0.74	1.18	0.93

Table 1. Linear regression results for eHiTS Score and binding affinity for the docked ligands.

This is an outstanding correlation for the docked ligands, that offers high level of confidence in the ranking of molecules. For ligands for which the top rank pose was under 2 Å RMSD compared to the crystallographic conformation (70% of the cases) the correlation was slightly improved as shown in Table 2.

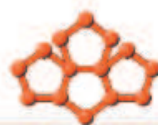
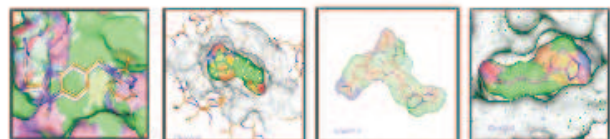
Scoring function	$R_p$	SD	ME
eHiTS-2009	0.76	1.12	0.88

Table 2. Linear regression results for eHiTS Score and binding affinity for top rank poses under 2 Å RMSD.

We further scored the crystallographic conformations of the ligands using eHiTS 2009 score utility. By neutralizing the docking aspect, we could compare the performance of eHiTS 2009 with other scoring functions. The results of this experiment are shown in Table 3.

Scoring function	$R_p$	SD	ME
eHiTS-2009	0.59	1.36	1.08

Table 3. Linear regression results for eHiTS Score and binding affinity for the X-ray poses of the ligands



SimBioSys®

## Technical Note

Table 4, cites statistics of similar models for the 2002 version of the refined set taken from references [3,4]. A direct comparison cannot be made given the different data set. Nevertheless, eHiTS is obviously among the top performing scoring functions for binding affinity correlation.

Scoring function	N	$R_P$	SD	ME
eHiTS 2009	800	0.64	1.34	1.06
X-Score::HMScore	800	0.57	1.82	1.42
DrugScore::Pair	800	0.47	1.94	1.51
AIScore+XFurcate	799	0.46	1.96	1.54
FlexX Score(Opt)	799	0.43	1.99	1.55
GOLD::ChemScore	762	0.45	1.96	1.52
GOLD::GoldScore	772	0.37	2.06	1.63

eHiTS 2009 is therefore an extremely powerful tool for docking and scoring potentially active molecules in a way that is well correlated with their binding affinity. eHiTS can thus reliably direct rational drug discovery efforts as a means for screening and prioritizing compounds for lab experiments.

### Data

The PDDBind data-set has been made available by the authors of references [1] and [2] at

<http://www.pdbbind.org/>

The refined set of PDDBind 2008 release consists of 1401 split complexes, covering a vast spectrum of protein families.