# Supplementary Information: A D3R Prospective Evaluation of Machine Learning for Protein-Ligand Scoring

August 25, 2016

## 1 Ligand-Based Regression Protocol

The CHEMBL3880 HSP90-alpha target dataset (`bioactivity-15_21_16_49.txt`) was downloaded from `https://www.ebi.ac.uk/chembl/`. The data was then processed to extract compounds with valid IC50 values using Python:

```
import pandas as pd
hsp = pd.read_csv('bioactivity-15_21_16_49.txt',sep='\t')
smi = hsp[(hsp.STANDARD_TYPE == 'IC50') & (hsp.RELATION == '=') &
        (hsp.STANDARD_UNITS == 'nM') & (hsp.PCHEMBL_VALUE > 0)]
        .loc[:,['CANONICAL_SMILES','PCHEMBL_VALUE']]
smi.to_csv('hsp90.smi',sep='\t',index=False,header=False)
```

Salts were then removed from `hsp90.smi` by extracting only the largest connected component. Scripts from `https://github.com/dkoes/qsar-tools` were then used to create models and make predictions.

### 1.1 RDKit

```
outputfingerprints.py hsp90.smi -o hsp90_rdkit_fp.gz --rdkit
trainlinearmodel.py -o hsp90_rdkit.model hsp90_rdkit_fp.gz  --maxiter 10000 --elastic
outputfingerprints.py --rdkit hsp90_test.smi -o hsp90_test_rdkit_fp.gz
../applylinearmodel.py hsp90_rdkit.model hsp90_test_rdkit_fp.gz > rdkit.out
join hsp90_test.smi rdkit.out | awk '{print $2,$3}' > LigandScores-1.csv
```

### 1.2 SMARTS

```
createsmartsdescriptors.py hsp90.smi -o hsp90.smarts
outputfingerprints.py hsp90.smi -o hsp90_smarts_fp.gz --smarts --smartsfile hsp90.smarts

trainlinearmodel.py -o hsp90_smarts.model hsp90_smarts_fp.gz  --elastic
outputfingerprints.py hsp90_test.smi --smartsfile hsp90.smarts -o hsp90_test_fp.gz --smarts
applylinearmodel.py hsp90_smarts.model hsp90_test_fp.gz > smarts.out
join hsp90_test.smi smarts.out | awk '{print $2,$3}' > LigandScores-2.csv
```

### 1.3 ECFP6

```
outputfingerprints.py hsp90.smi --ecfp6 -o hsp90_ecfp6_fp.gz
trainlinearmodel.py -o hsp90_ecfp6.model hsp90_ecfp6_fp.gz --elastic
outputfingerprints.py --ecfp6 hsp90_test.smi -o hsp90_test_ecfp6_fp.gz
applylinearmodel.py hsp90_ecfp6.model  hsp90_test_ecfp6_fp.gz > ecfp.out
join hsp90_test.smi ecfp.out | awk '{print $2,$3}' > LigandScores-3.csv
```

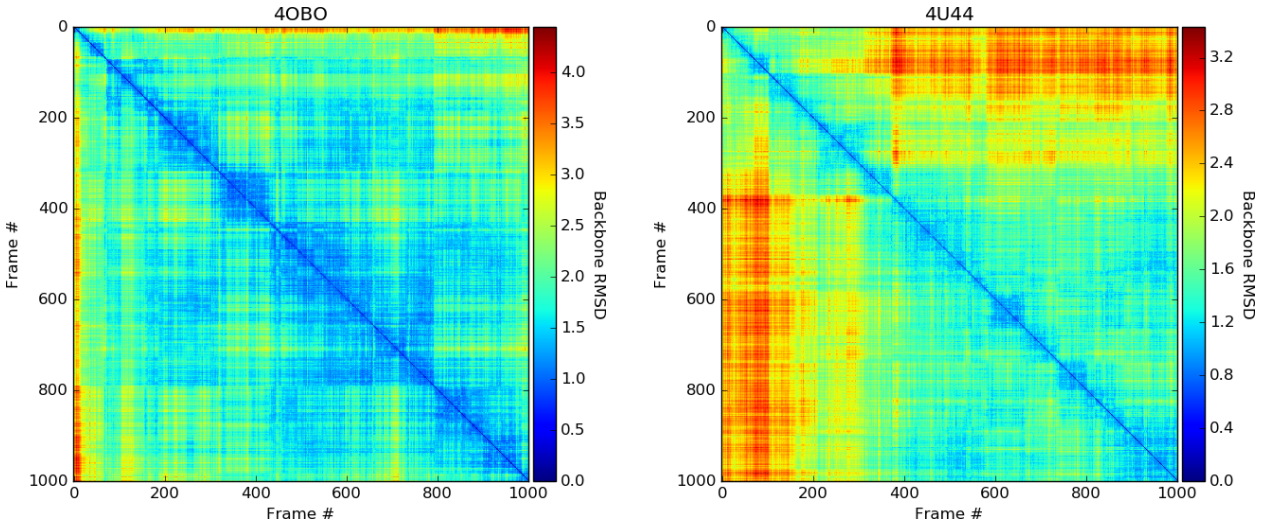


Figure S1: Heatmaps of backbone RMSDs between frames within 100ns trajectories of MAP4K4 molecular dynamics simulations.