Machine Learning to Predict Experimental Protein-Ligand Complexes

Hyunji Kim[1], Sarah Walworth[2], Kenny Merz[3], Jun Pei[3], Lin Song[3], Zheng Zheng[3]
[1]George Washington University,[2]University of Colorado Boulder, [3]Michigan State University

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

OBJECTIVE

Traditional scoring methods to determine correct poses for protein-ligand binding are generally **around 60% accurate**. Our goal was to use **random forest machine learning** to optimize **the ability to predict ligand poses** that are close to the native crystal structure of the protein-ligand complex. One major application of our method is **drug design**. It will allow "designers" to find molecules that could dock similarly to the native crystal structure.

METHOD

1. Data Generation:
Given **766 protein-ligand complexes**, we generated ligand decoys (up to 100 per protein) using **Schrodinger Glide software.**

2. GARF Potential Function:
We considered approximate effects on energy using **the GARF pairwise interatomic potential function.**

3. Random Forest (RF)
We generated our RF model using the Scikit-Lean tool. **5-fold cross validation** was applied to prevent overfitting. For every simulation, we randomly selected 70% of the data as the training set and 30% as the test set.

Classification:

$$Rank_{native} < Rank_{ligand_{decoy}} \Longrightarrow Class\ 0$$
$$Rank_{native} > Rank_{ligand_{decoy}} \Longrightarrow Class\ 1$$

4. Scoring
We used the Cambridge Crystallographic Data Center (CCDC) GOLD protein-ligand docking software to generate the **Astex Statistical Potential(ASP)** and **Chemscore** scoring functions to validate our RF model.

5. Post-Processing
We performed **17 grid searches** to narrow down the RF parameters. From the grid search, we **selected the best 6 parameters** and ran 12 independent simulations for each of the 6 parameter combinations **to identify the best parameter** for our RF model.

RANDOM FOREST MODEL

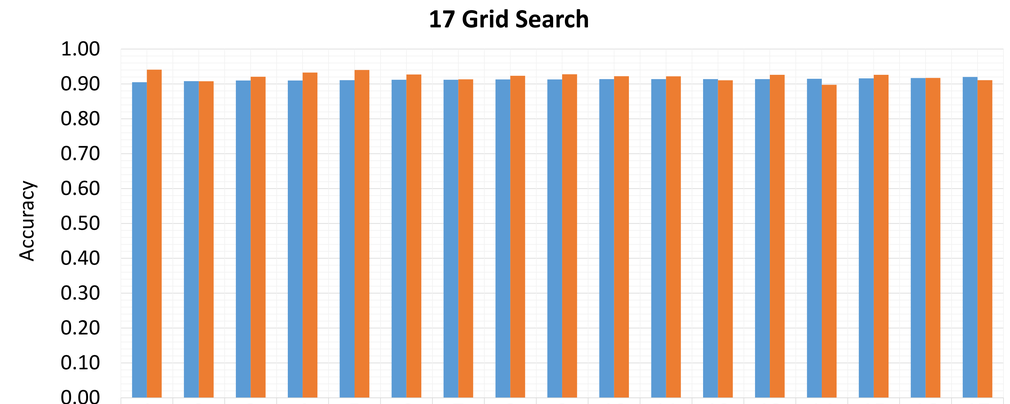

Figure 1: Results of 17 independent grid searches to narrow the RF parameters

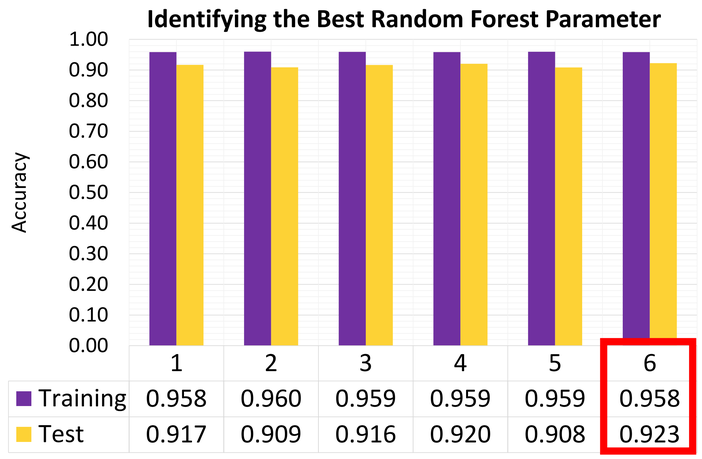

Figure 2: The averaged training and test accuracies for the six best RF parameters; for each parameter combination, we ran 12 independent RF simulations

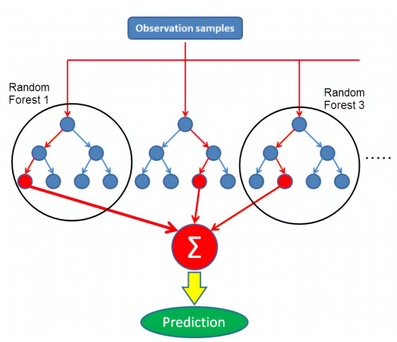

Figure 3: Graphical representation of the Random forest model [1]

Finalized RF Model	
N_estimators	1000
Max_depth	10
Min_sample split	5
Min_samples leaf	2

Table 1: Final RF Parameter (Choose Parameter 6 from figure 2)

VALIDATION

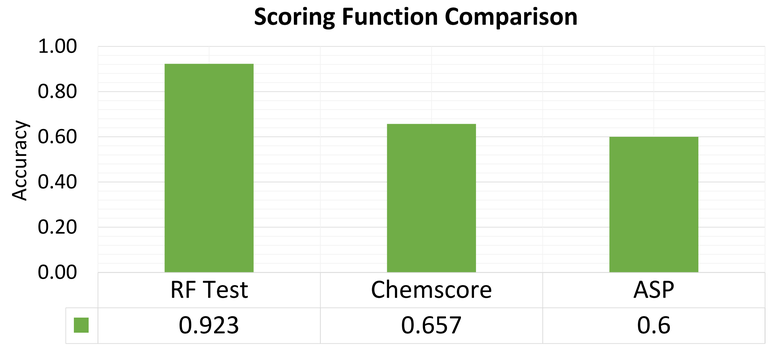

Figure 3: Comparison of RF results against the results of two traditional scoring functions (Chemscore and ASP)

Chemscore:
- **Empirical scoring function**: Regression based with coefficients based on experimental data, which accounted for physical factors that affect docking.

Astex Statistical Potential (ASP):
- **Atom to atom potential function** using the Worldwide Protein Data Bank.
- Considered frequency and potentials: Expected number of interactions of atoms in a defined radius.

CONCLUSION

Overall Results:
- Our results have shown that our random forest machine learning model is **significantly more accurate** in predicting ligand poses similar to the native crystal structure of a protein-ligand complex than two traditional scoring functions.

Future Work:
- For further validation, we plan to **test our model with a larger data set** using the sets of decoys that have been generated in **the Database of Useful Decoys: Enhanced (DUD-E).**
- We also plan to use **the GARF scoring function** as an accuracy comparison.

REFERENCES

[1] T. (2018, May 01). Machine Learning for Beginners, Part 10: Random Forest. Retrieved from https://thedatalass.com/2018/04/17/random-forest/

The Cambridge Crystallographic Data Centre

Advanced Computational Research Experience for Students

We acknowledge support from the MSU ACRES REU program, which is supported by the National Science Foundation through grant ACI-1560168.