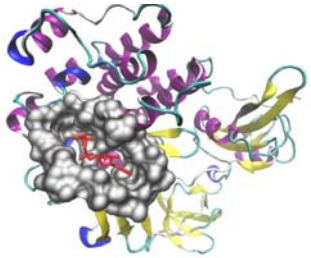
How Low Can You Go? Feature Selection for Drug Discovery

Derek Jones*, Sally R. Ellingson*, W.A. de Jong

*University of Kentucky

Background and Motivation

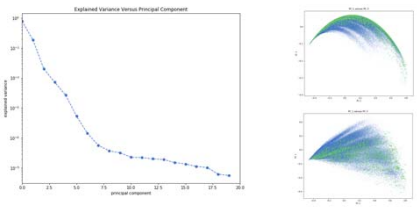
- The cost of bringing a drug to market depends on how quickly a candidate drug can be “discovered” and evaluated to ensure safety and effectiveness
- In this work we develop a method for predicting whether a given drug and protein compound will “bind”.
- Our aim is to select a set of features to predict drug-protein interactions



This study focuses on kinases. Kinase inhibitors are the largest class of new cancer therapies. Selective inhibition is difficult due to high sequence similarity, leading to off-target interactions and side-effects. Pictured here human c-SRC.

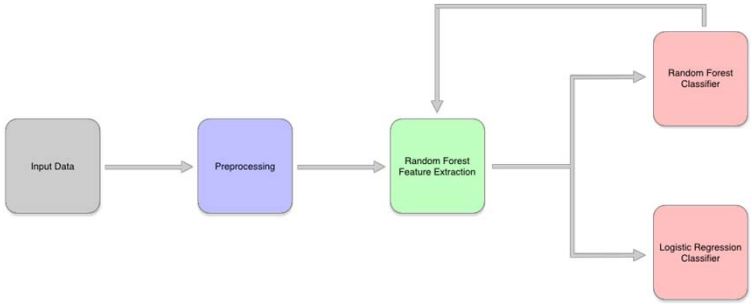
Dataset

- **Our dataset consists of 361,786 protein-drug molecule combinations** from the Directory of Useful Decoys Enhanced [4] subset of kinases which includes both known active compounds and generated decoys for 26 kinases. We collected the following features for our dataset:
  - Binding features: Vina MPI [2]
  - Drug features: Dragon [1]
  - Protein features: ExPasy [6], Porter,PaleAle 4.0 [5], & PROFEAT-Protein Feature Server [7]
  - Pocket features* [8]
- **1:50 ratio of positive to negative training examples**
- **5432 features before selection pipeline, reduced to a set of 1260** which are examined using PCA.



*to be included

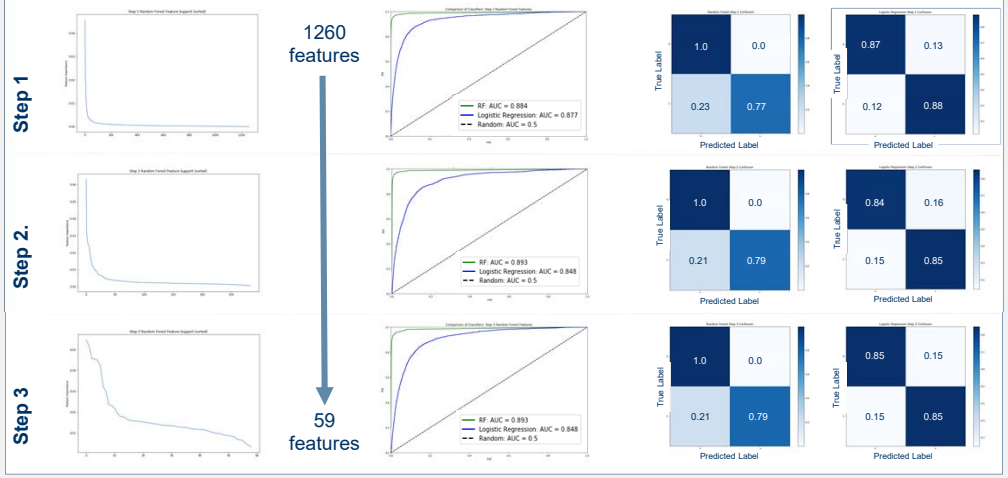
Methods



1. **Preprocessing:** Impute data using mean for each feature, then normalize each feature to unit length
2. **Random Forest Feature Extraction:** train a random forest, using randomized grid search. Using the feature importances of the optimal random forest classifier, create a reduced feature set from the features with above mean importance.
3. **Create 80/20 training and testing stratified split** of the data using only the “relevant” features
4. **Train** classification models on the reduced feature set, using randomized grid search to select the optimal model parameters.
5. **Test** classification models on the reduced feature set
6. **Repeat** until a minimal set of features are selected

Results

Given the initial set of 5432 features, we are able to reduce this set by 2 orders of magnitude while retaining nearly identical performance on the classification task. We evaluate a random forest and logistic regression on each reduced set.



Results cont..

Table 1: Random Forest Performance

Reduction	N Features	Precision	Recall	F1-Score	Positive Precision	Positive F1
Step 1	1260	0.99	0.99	0.99	0.99	0.87
Step 2	284	0.99	0.99	0.99	0.94	0.86
Step 3	59	0.99	0.99	0.99	0.93	0.85
Step 4	15	0.99	0.99	0.99	0.68	0.74

Table 2: Logistic Regression Performance

Reduction	N Features	Precision	Recall	F1-Score	Positive Precision	Positive F1
Step 1	1260	0.97	0.87	0.91	0.16	0.26
Step 2	284	0.97	0.84	0.89	0.12	0.22
Step 3	59	0.97	0.85	0.90	0.13	0.22
Step 4	15	0.97	0.79	0.86	0.10	0.17

Conclusions and Future Work

- We are able to significantly reduce the feature set and identify the important properties of the interaction to make accurate predictions
- This work helps lay the foundation for future work that will ask more specific questions regarding protein-drug molecule interactions
- Can we expand our model to include multiple protein binding pockets to understand more complex interactions?
- Can we develop an effective method to predict adverse drug reactions based upon a drug molecule binding to multiple proteins?
- Can we use secondary structure information about the protein to improve our results?

Acknowledgements

This research used computational resources at the University of Kentucky’s Center for Computational Sciences and the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This work was supported by the National Institutes of Health (NIH) National Center for Advancing Translational Science grant KL2TR000116 and 1KL2TR001996-01.This work was supported by the Director, Office of Science, Office of Advanced Scientific Computing Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

1. Kode srl, Dragon (software for molecular descriptor calculation) version 7.0.6, 2016, https://chm.kode-solutions.net
2. Ellingson, Sally R., Jeremy C. Smith, and Jerome Baudry. 2013. “VinaMPI: Facilitating Multiple Receptor High-Throughput Virtual Docking on High-Performance Computers.” *Journal of Computational Chemistry* 34 (25). Wiley Online Library: 2212–21.
3. Jamali, Ali Akbar, Reza Ferdousi, Saeed Razzaghi, Jiuyong Li, Reza Safdari, and Esmaeil Ebrahimie. 2016. “DrugMiner: Comparative Analysis of Machine Learning Algorithms for Prediction of Potential Druggable Proteins.” *Drug Discovery Today* 21 (5): 718–24.
4. Mysinger, Michael M., Michael Carchia, John J. Irwin, and Brian K. Shoichet. 2012. “Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking.” *Journal of Medicinal Chemistry* 55 (14): 6582–94.
5. Mirabello, Claudio, and Gianluca Pollastri. 2013. “Porter, PaleAle 4.0: High-Accuracy Prediction of Protein Secondary Structure and Relative Solvent Accessibility.” *Bioinformatics* 29 (16): 2056–58.
6. Gasteiger, Elisabeth, Christine Hoogland, Alexandre Gattiker, S ’everine Duvaud, Marc R. Wilkins, Ron D. Appel, and Amos Bairoch. 2005. “Protein Identification and Analysis Tools on the ExPASy Server.” In *The Proteomics Protocols Handbook*, edited by John M. Walker, 571–607. Totowa, NJ: Humana Press.
7. Zhang, Peng, Lin Tao, Xian Zeng, Chu Qin, Shangying Chen, Feng Zhu, Zerong Li, Yuyang Jiang, Weiping Chen, and Yu-Zong Chen. 2016. “A Protein Network Descriptor Server and Its Use in Studying Protein, Disease, Metabolic and Drug Targeted Networks.” *Briefings in Bioinformatics*, August. doi:10.1093/bib/bbw071.
8. Kriv ́ak, Radoslav, and David Hoksza. 2015. “Improving Protein-Ligand Binding Site Prediction Accuracy by Classification of Inner Pocket Points Using Local Features.” *Journal of Cheminformatics* 7 (April): 12.