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

Methods
1. Preprocessing: Impule 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

Results
Our dataset consists of 361,786 protein-drug molecule combinations from the Directory of Useful Decoys (DUD). 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.

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 KL2TR001116 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