AMIA

INFORMATICS PROFESSIONALS. LEADING THE WAY.

Polypharmacology Within the Full Kinome: A Machine Learning Approach

Training Machines to Make Discoveries

S37

Derek Jones

University of Kentucky

Disclosure

I have no relevant relationships with commercial interests to disclose.

Goals

After participating in this session the learner should be better able to:

- Understand the significant costs associated with Drug Discovery and the importance of developing efficient computational tools
- Identify an instance of a machine learning problem in which class imbalance can hide biases in the model through the lens of a single metric
- Understand how to use a data-driven feature selection approach

Motivation: Drug Discovery is Challenging

- There have been several attempts to quantify the price of drug discovery in recent years:
- In 2010 the estimated cost of bringing a new drug to market was $1.8 billion.
- A late 2014 estimate rose to $2.6 billion dollars.

As Drug discovery becomes more expensive, it is also becoming more difficult

(Scannel et. al, Nat. Rev. Drug Discov.)

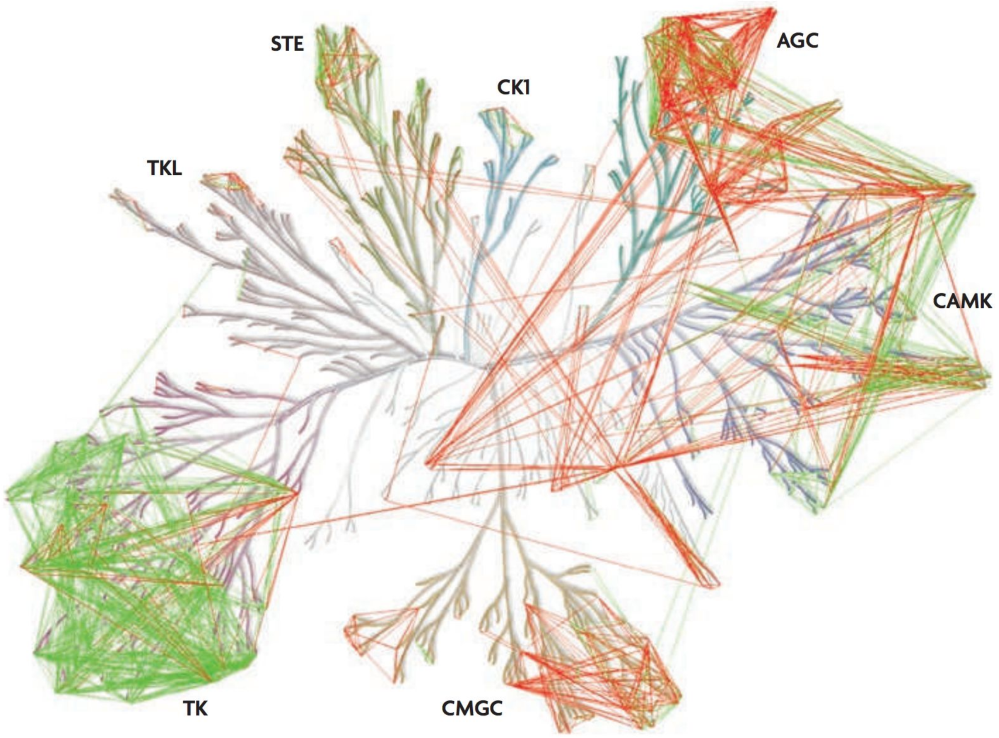
Solutions: Polypharmacology

- A one-to-many relationship between a drug molecule and possible target proteins.
- Establishing networks of how drugs interact with many proteins can offer insight into drug repurposing, side-effect prediction, and the development of more efficacious drugs.

What to proteins to target?

- Protein kinases regulate the majority of cellular pathways and signal transduction
- The deregulation of kinases has been implicated in many disease states, especially in cancers
- Due to the high similarity in sequence and structure between kinases, selectivity is a huge challenge for drug design

Selectivity in Kinases



- Pairs of kinases that were potently inhibited by a common inhibitor (green lines) were used to determine a sequence similarity cut-off to predict pairs of kinases that can be inhibited by a common inhibitor (red lines)

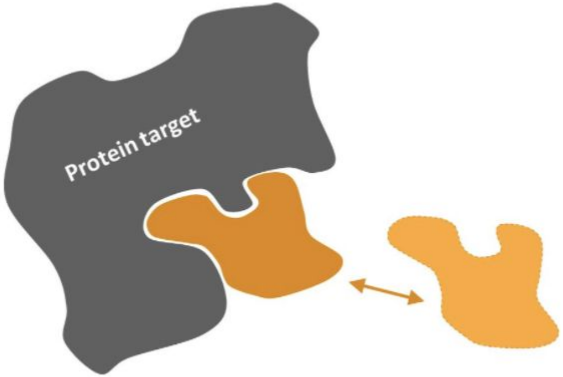
What’s a Computer Scientist to Do?

- Can we leverage big data to learn the likelihood of a drug binding to a protein with much greater accuracy than current simulation methods alone?
- Can we identify the features that are informative?
- Can we do this with a method that is robust to the class imbalance between actives (positive) and decoys (negative), that can process large sets of features?

What’s a Computer Scientist to Do?

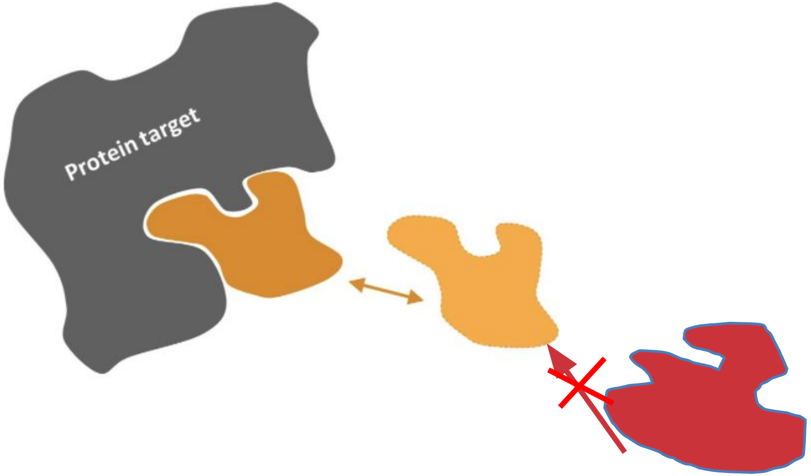
- Can we leverage big data to learn the likelihood of a drug binding to a protein with much greater accuracy than current simulation methods alone?
- Can we identify the features that are informative?
- Can we do this with a method that is robust to the class imbalance between actives (positive) and decoys (negative), that can process large sets of features?

What are we modeling?

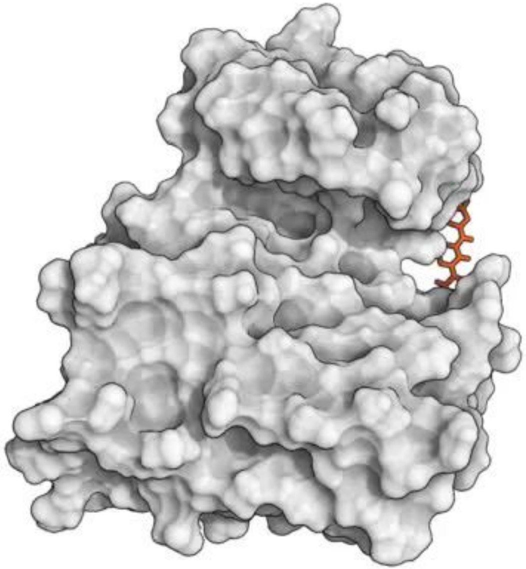


Protein-Drug Interaction:
Active
Binders

Protein-Drug non-Interaction:
Decoy
Non-Binders



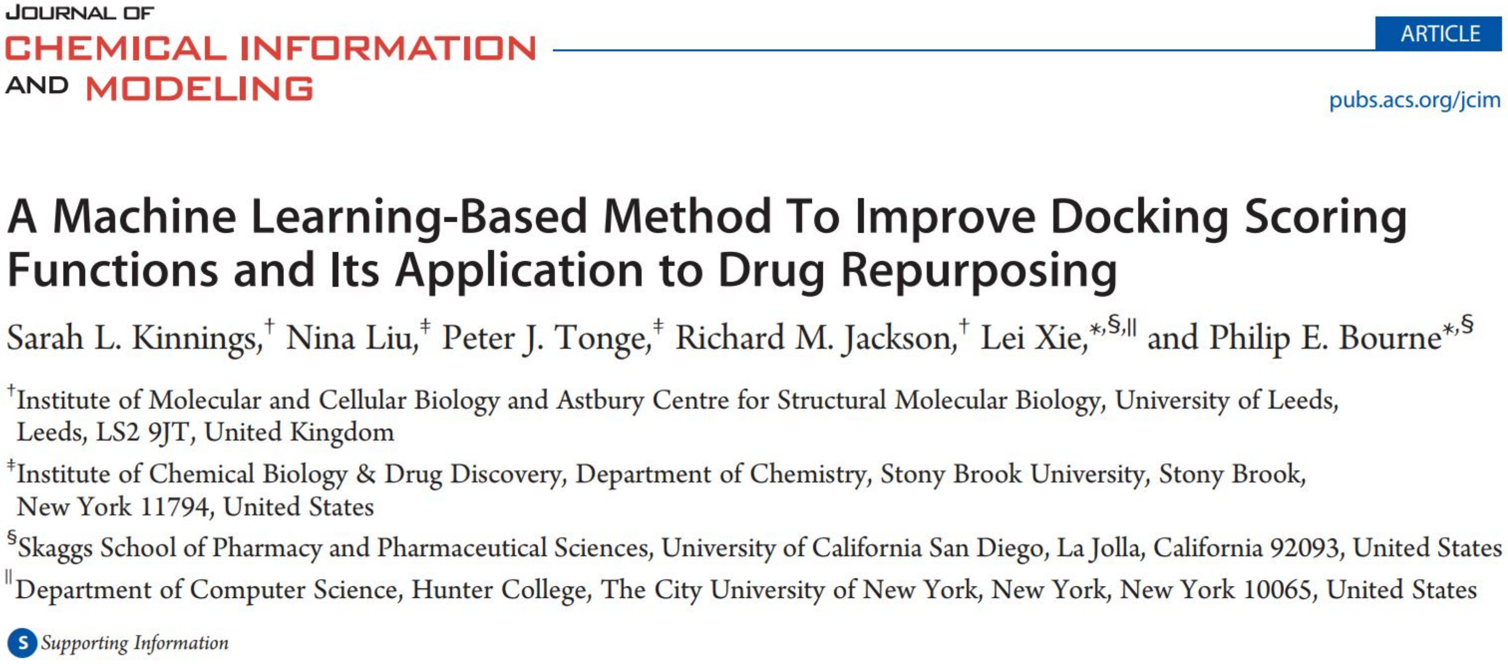
What are we modeling?



Binding of cancer drug dasatinib to target, Src kinase

How does a drug molecule find its target binding site?
Shan et al. (2011) JACS. (Anton)

Background: Molecular Docking Features



JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

ARTICLE

pubs.acs.org/jcim

A Machine Learning-Based Method To Improve Docking Scoring Functions and Its Application to Drug Repurposing

Sarah L. Kinnings,[†] Nina Liu,[‡] Peter J. Tonge,[‡] Richard M. Jackson,[†] Lei Xie,[*,§,||] and Philip E. Bourne[*,§]

[†]Institute of Molecular and Cellular Biology and Astbury Centre for Structural Molecular Biology, University of Leeds, Leeds, LS2 9JT, United Kingdom

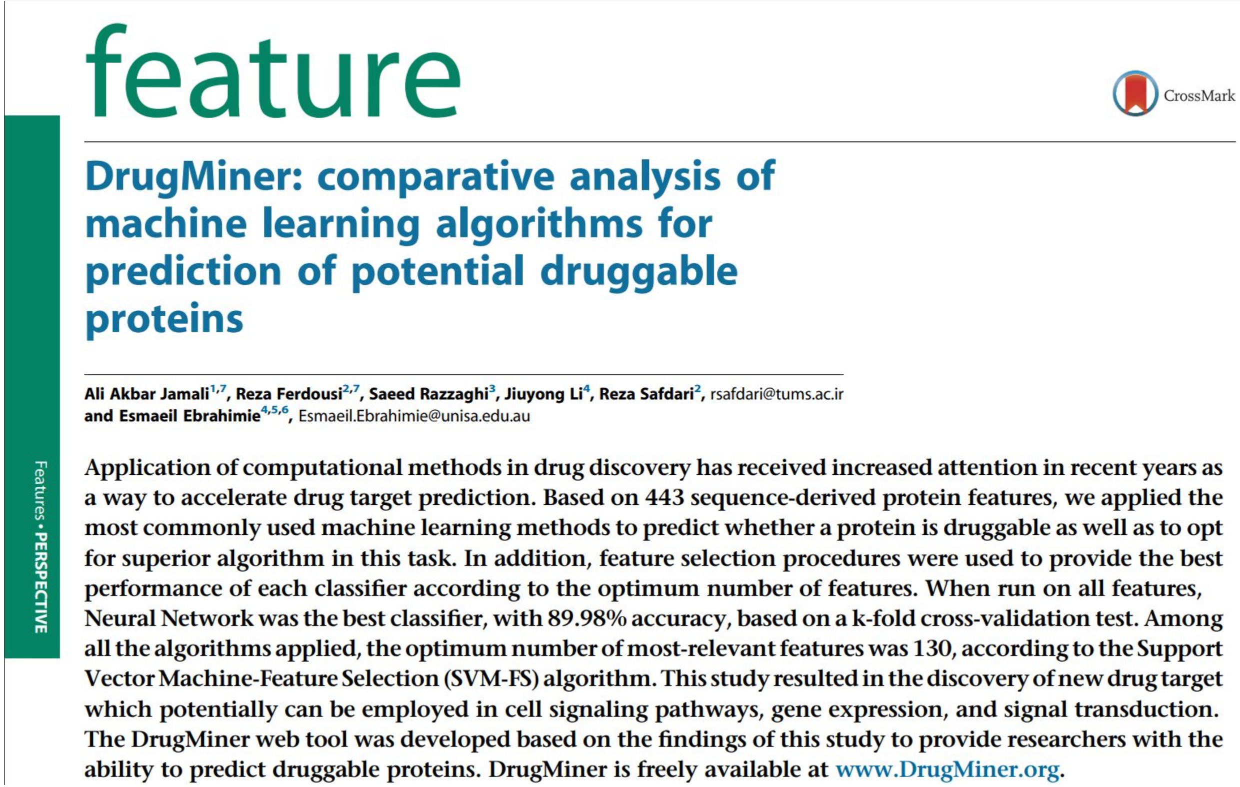
[‡]Institute of Chemical Biology & Drug Discovery, Department of Chemistry, Stony Brook University, Stony Brook, New York 11794, United States

[§]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, California 92093, United States

[||]Department of Computer Science, Hunter College, The City University of New York, New York, New York 10065, United States

S *Supporting Information*

Background: Protein Features



feature

Features • PERSPECTIVE

DrugMiner: comparative analysis of machine learning algorithms for prediction of potential druggable proteins

Ali Akbar Jamali[1,7], Reza Ferdousi[2,7], Saeed Razzaghi[3], Jiuyong Li[4], Reza Safdari[2], rsafdari@tums.ac.ir
and Esmaeil Ebrahimie[4,5,6], Esmaeil.Ebrahimie@unisa.edu.au

Application of computational methods in drug discovery has received increased attention in recent years as a way to accelerate drug target prediction. Based on 443 sequence-derived protein features, we applied the most commonly used machine learning methods to predict whether a protein is druggable as well as to opt for superior algorithm in this task. In addition, feature selection procedures were used to provide the best performance of each classifier according to the optimum number of features. When run on all features, Neural Network was the best classifier, with 89.98% accuracy, based on a k-fold cross-validation test. Among all the algorithms applied, the optimum number of most-relevant features was 130, according to the Support Vector Machine-Feature Selection (SVM-FS) algorithm. This study resulted in the discovery of new drug target which potentially can be employed in cell signaling pathways, gene expression, and signal transduction. The DrugMiner web tool was developed based on the findings of this study to provide researchers with the ability to predict druggable proteins. DrugMiner is freely available at www.DrugMiner.org.

Feature Representation

- Bourne et. al., use the individual components of the docking scoring function as features in a single protein druggability model
- DrugMiner makes exclusive use of protein sequence features to classify druggable proteins in a multi-protein model
- Our methods combine these feature representations along with molecular descriptors and binding pocket features in a multi-protein model

Feature Representation

- Bourne et. al., use the individual components of the docking scoring function as features in a single protein druggability model
- DrugMiner makes exclusive use of protein sequence features to classify druggable proteins in a multi-protein model
- Our methods combine these feature representations along with molecular descriptors and binding pocket features in a multi-protein model
- In total have 5,410 features among all groups

DUD•E

A Database of Useful Decoys: Enhanced

An enhanced and rebuilt version of DUD, a directory of useful decoys. DUD-E is designed to help benchmark molecular docking programs by providing challenging decoys. It contains:

- 22,886 active compounds and their affinities against 102 targets, an average of 224 ligands per target
- 50 decoys for each active having similar physico-chemical properties but dissimilar 2-D topology.
- Kinase subset of DUD-e (26 kinases)

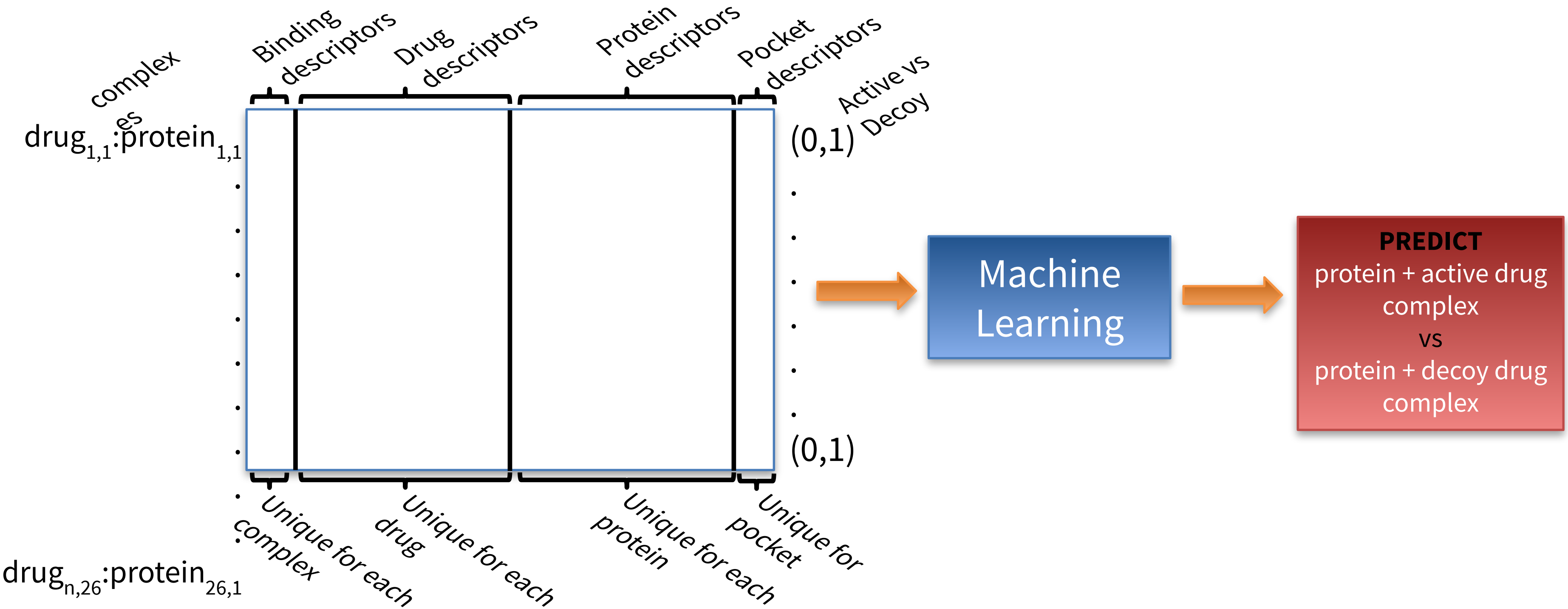
DUD-E is provided by the Shoichet Laboratory in the Department of Pharmaceutical Chemistry at the University of California, San Francisco (UCSF)

http://dude.docking.org/

Dataset Description

- **Binding Features:** Collected from Molecular Docking calculation scoring function
- **Drug Features:** Computed using Dragon 7, used all features except for 3D descriptors
- **Protein Features:** All derived from sequence, all features in DrugMiner
- **Pocket Features:** PRANK, software used to predict and rank binding sites
- Final dataset includes 361,786 training examples, total of 5,410 features

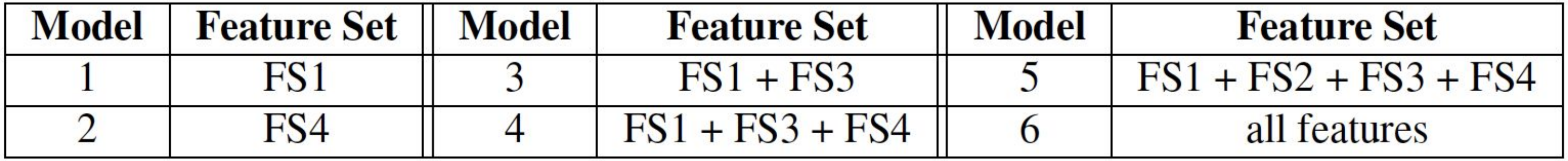
Dataset Description



Modeling What's Important

- Feature Set 1 **(FS1)**: This set is selected using the entire dataset and using the active or decoy binary labels.
- Feature Set 2 **(FS2)**: This set is selected using only protein and pocket features and using the kinase as a label.
- Feature Set 3 **(FS3)**: This set is selected using the drug features with the kinase as a label.
- Feature Set 4 **(FS4)**: This set contains all docking features

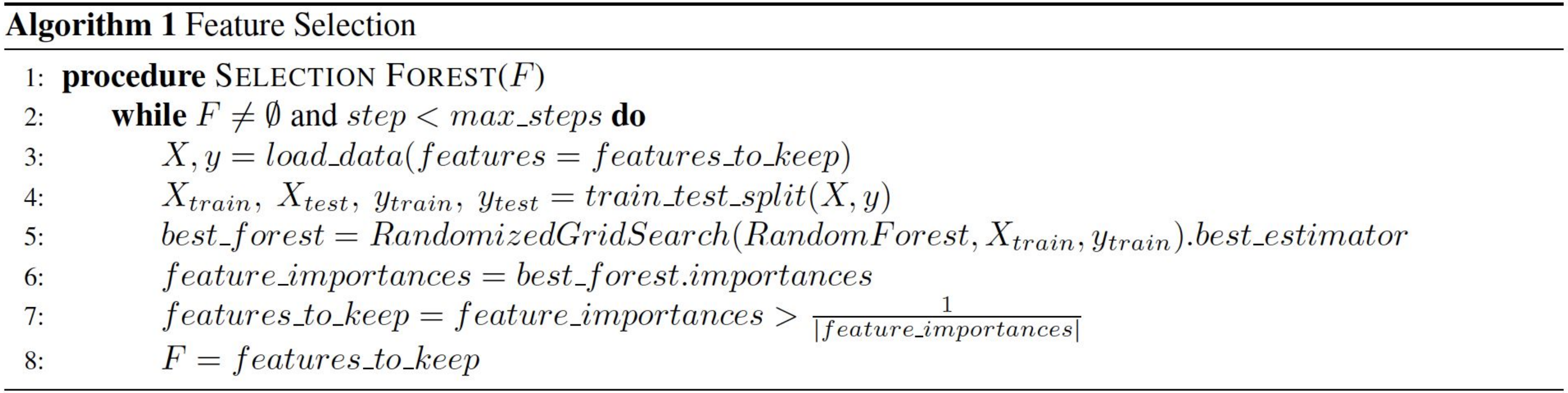
Models



Model	Feature Set	Model	Feature Set	Model	Feature Set
1	FS1	3	FS1 + FS3	5	FS1 + FS2 + FS3 + FS4
2	FS4	4	FS1 + FS3 + FS4	6	all features

Use a number of choices for initial feature sets in order to extract important features w.r.t each group and to assess their predictive performance

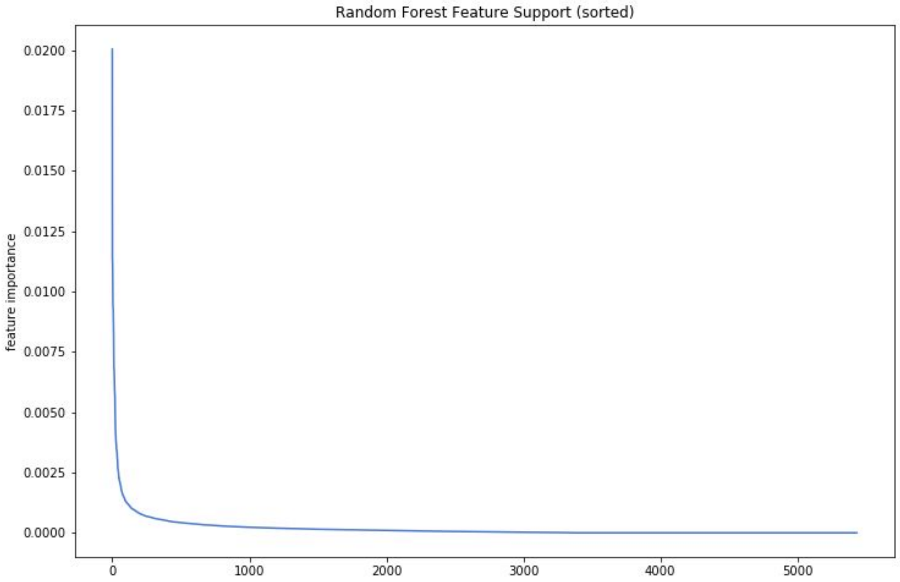
Feature Selection



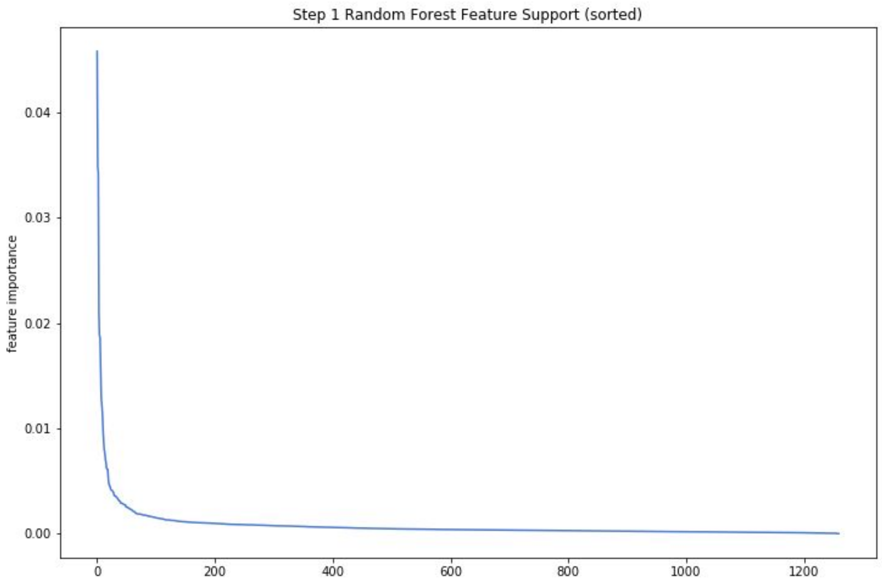
Algorithm 1 Feature Selection

```
procedure SELECTION FOREST($F$)
    while $F \neq \emptyset$ and $step < max\_steps$ do
        $X, y = load\_data(features = features\_to\_keep)$
        $X_{train},\ X_{test},\ y_{train},\ y_{test} = train\_test\_split(X, y)$
        $best\_forest = RandomizedGridSearch(RandomForest, X_{train}, y_{train}).best\_estimator$
        $feature\_importances = best\_forest.importances$
        $features\_to\_keep = feature\_importances > \frac{1}{|feature\_importances|}$
        $F = features\_to\_keep$
```

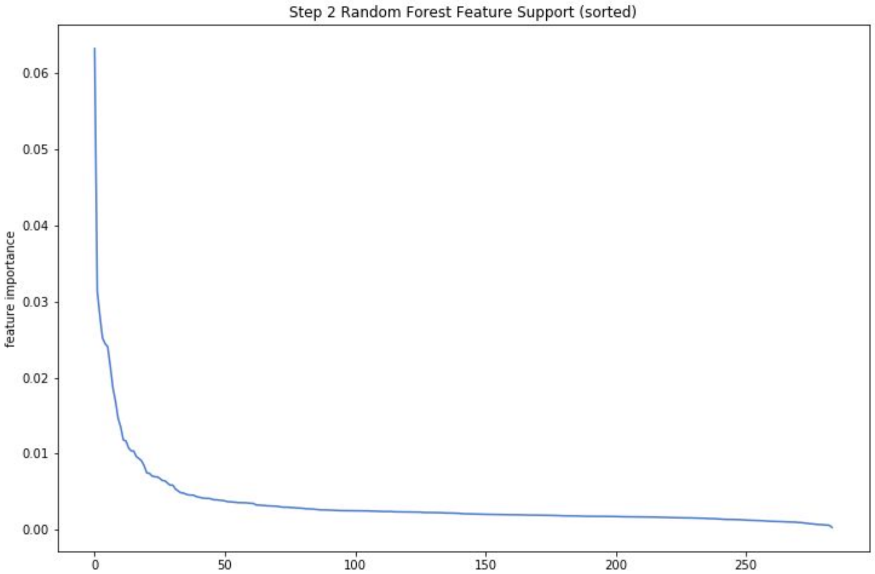
Feature Selection



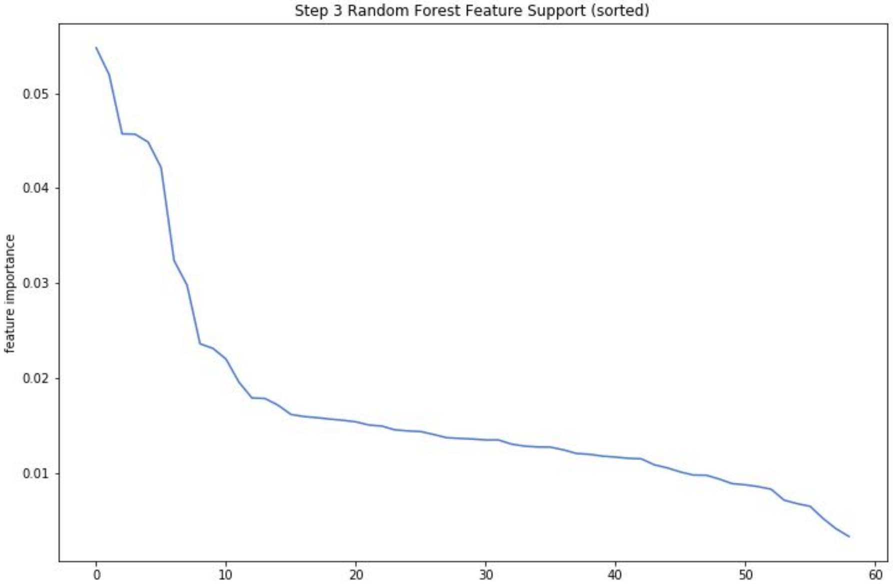
Feature Selection



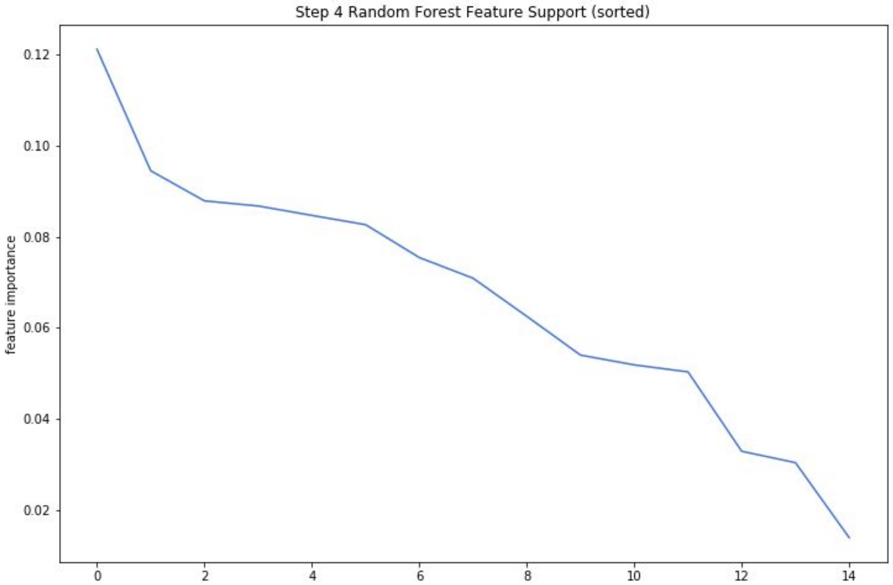
Feature Selection



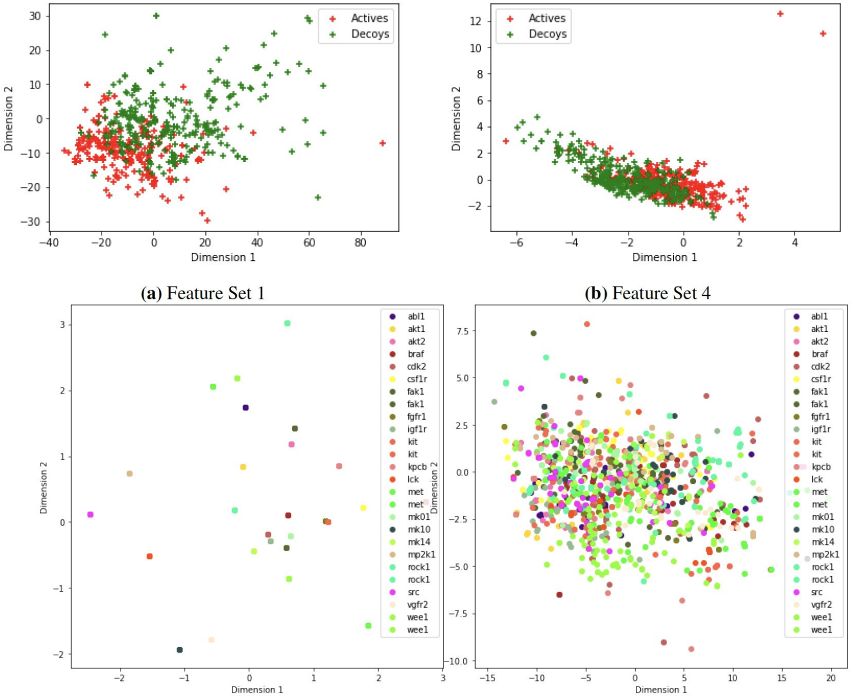
Feature Selection



Feature Selection



Feature Set Visualization



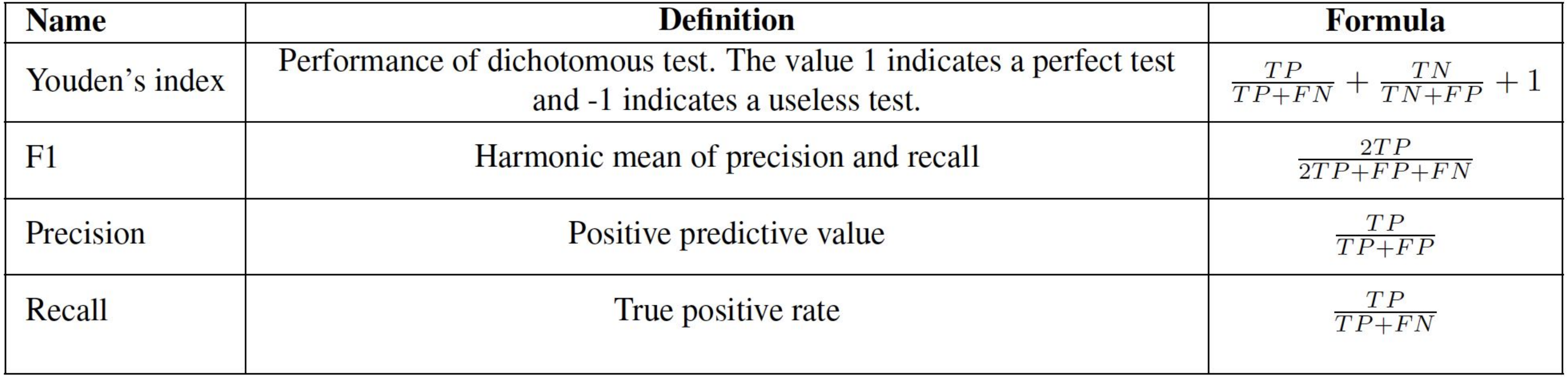
(a) Feature Set 1

(b) Feature Set 4

(c) Feature Set 2

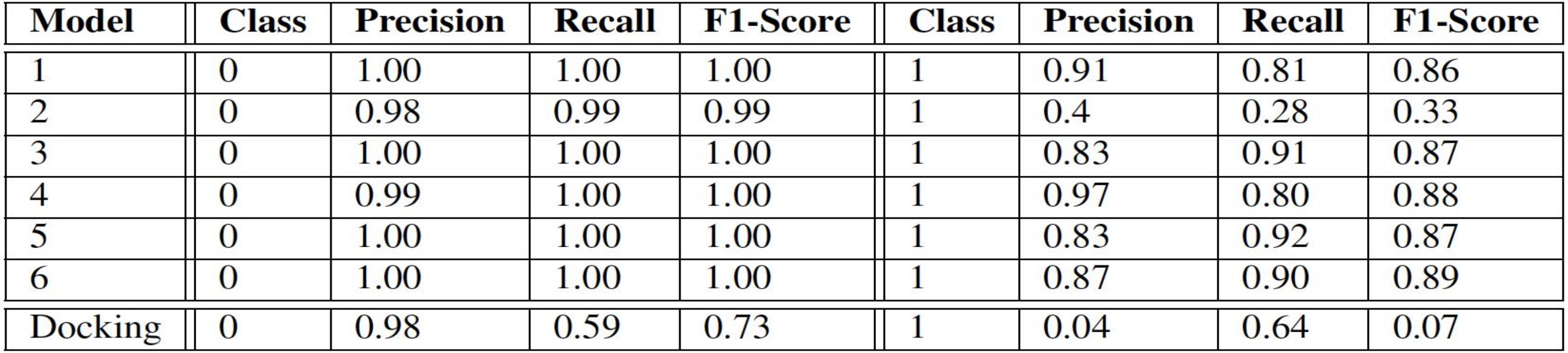
(d) Feature Set 3

Evaluation Metrics

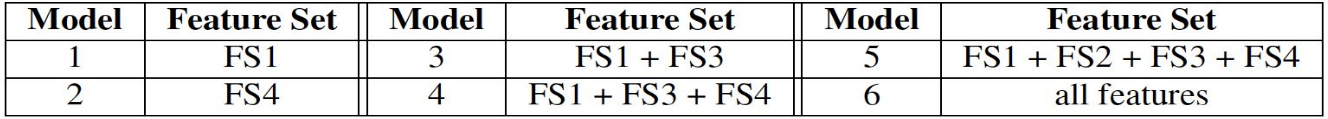


Name	Definition	Formula
Youden's index	Performance of dichotomous test. The value 1 indicates a perfect test and -1 indicates a useless test.	$\frac{TP}{TP+FN} + \frac{TN}{TN+FP} + 1$
F1	Harmonic mean of precision and recall	$\frac{2TP}{2TP+FP+FN}$
Precision	Positive predictive value	$\frac{TP}{TP+FP}$
Recall	True positive rate	$\frac{TP}{TP+FN}$

Classification Performance



Model	Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score
1	0	1.00	1.00	1.00	1	0.91	0.81	0.86
2	0	0.98	0.99	0.99	1	0.4	0.28	0.33
3	0	1.00	1.00	1.00	1	0.83	0.91	0.87
4	0	0.99	1.00	1.00	1	0.97	0.80	0.88
5	0	1.00	1.00	1.00	1	0.83	0.92	0.87
6	0	1.00	1.00	1.00	1	0.87	0.90	0.89
Docking	0	0.98	0.59	0.73	1	0.04	0.64	0.07



Model	Feature Set	Model	Feature Set	Model	Feature Set
1	FS1	3	FS1 + FS3	5	FS1 + FS2 + FS3 + FS4
2	FS4	4	FS1 + FS3 + FS4	6	all features

Conclusions

- We found that the features selected from our method greatly increased the performance of binding predictions

- Our methods achieve the goal of robustness to class imbalance and are able to extract the important features for the task in a data-driven way

- We observe near 60% increase in the success rate for discovering active compounds over molecular docking

Acknowledgements

- Cancer Research Informatics Shared Resource Facility of the University of Kentucky Markey Cancer Center (P30CA177558).
- University of Kentucky 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
- National Institutes of Health (NIH) National Center for Advancing Translational Science grant KL2TR000116 and1KL2TR001996-01
- We would like to acknowledge the assistance of Radoslav Krivak and David Hoksza in using the p2rank software.

Acknowledgements

UNIVERSITY OF KENTUCKY
CCTS
CENTER FOR CLINICAL AND TRANSLATIONAL SCIENCE

UKHealthCare®
Markey Cancer Center

BERKELEY LAB
Lawrence Berkeley National Laboratory

Thank you!

Email me at:
derek.jones4@uky.edu

AMIA is the professional home for more than 5,400 informatics professionals, representing frontline clinicians, researchers, public health experts and educators who bring meaning to data, manage information and generate new knowledge across the research and healthcare enterprise.

AMIA

INFORMATICS PROFESSIONALS. LEADING THE WAY.

f @AMIAInformatics

@AMIAinformatics

in Official Group of AMIA

YouTube @AMIAInformatics

#WhyInformatics