A View Towards Target Gene Identification Through CRISPR KO Labelling

Introduction

Motivation: People want to avoid costly development of a modulator that is ineffective at treating the pathology of interest due to bad target identification

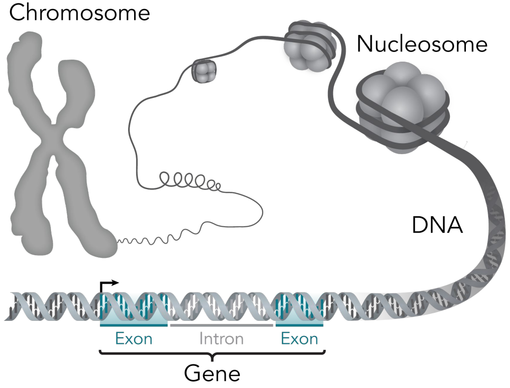
Goal: Target identification and validation (classify human genes into “targets” and “non-targets”)

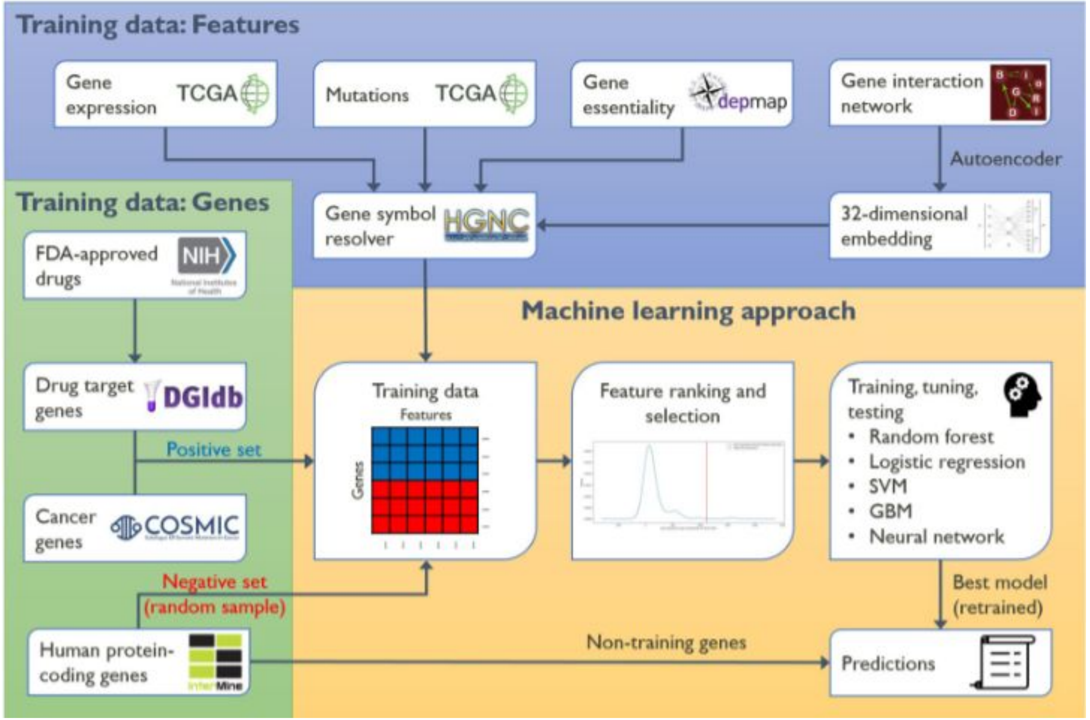
To predict novel therapeutic targets in oncology using computational intelligence methods

What they did in this paper: Different machine learning classifiers applied to the task of drug target classification for nine different human cancer type; predict on more than 15000 protein-coding genes

Method:

Data type: gene mutations and gene expression, essentiality data and features with a numerical embedding of the interaction network of protein-coding genes.
Six different machine learning classifiers: Random forest, artificial neural network, support vector machine, logistic regression, Linear Discriminant analysis (LDA) and KNN method.
What’ differ than other methods: disease-specific but cover a broad range of cancers





Training data: Features

Gene expression TCGA

Mutations TCGA

Gene essentiality depmap

Gene interaction network

Autoencoder

32-dimensional embedding

Gene symbol resolver HGNC

Training data: Genes

FDA-approved drugs NIH

Drug target genes DGIdb

Positive set

Cancer genes COSMIC

Negative set (random sample)

Human protein-coding genes InterMine

Machine learning approach

Training data

Features

Genes

Feature ranking and selection

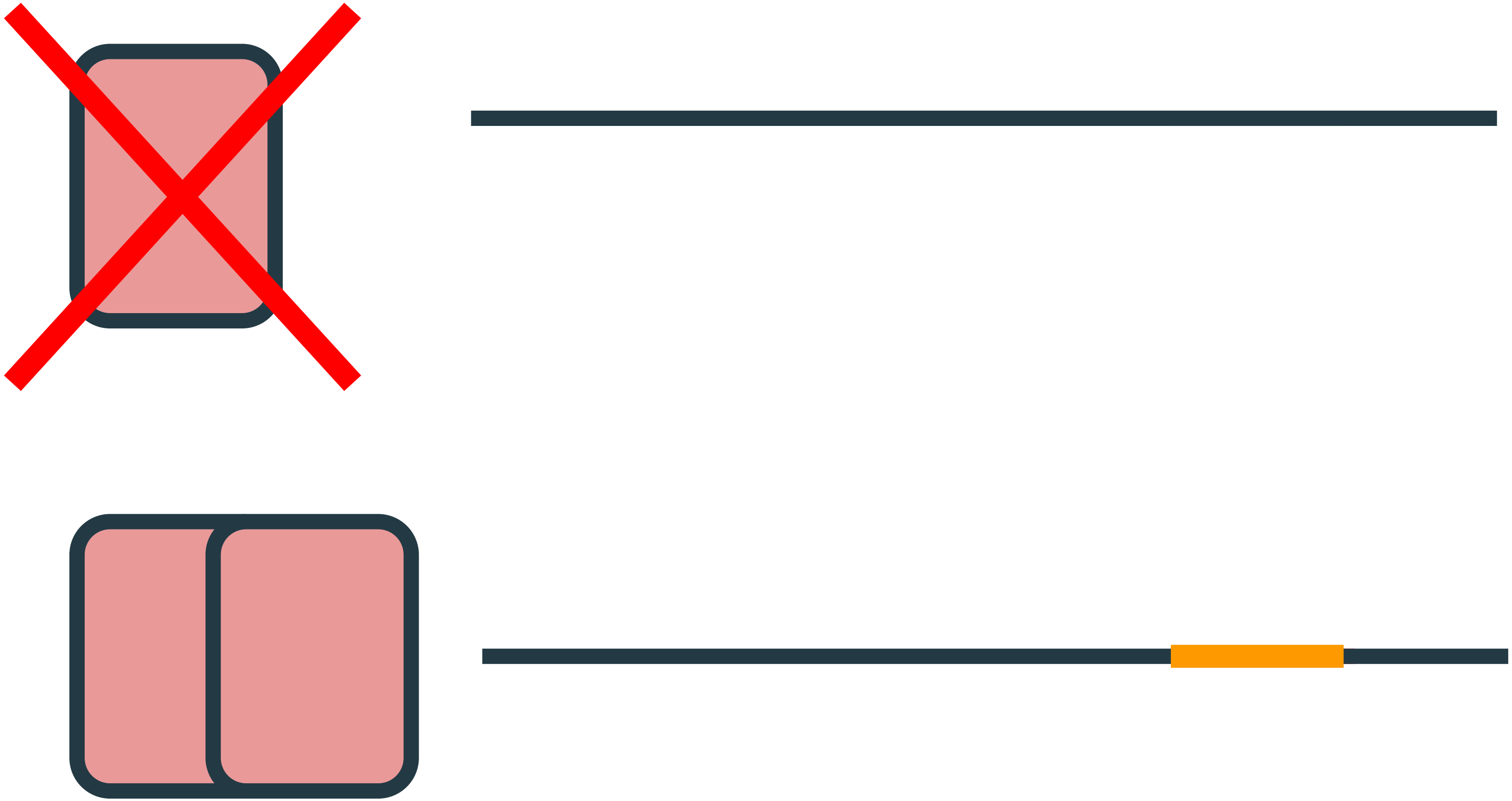
Training, tuning, testing

- Random forest
- Logistic regression
- SVM
- GBM
- Neural network

Best model (retrained)

Non-training genes

Predictions



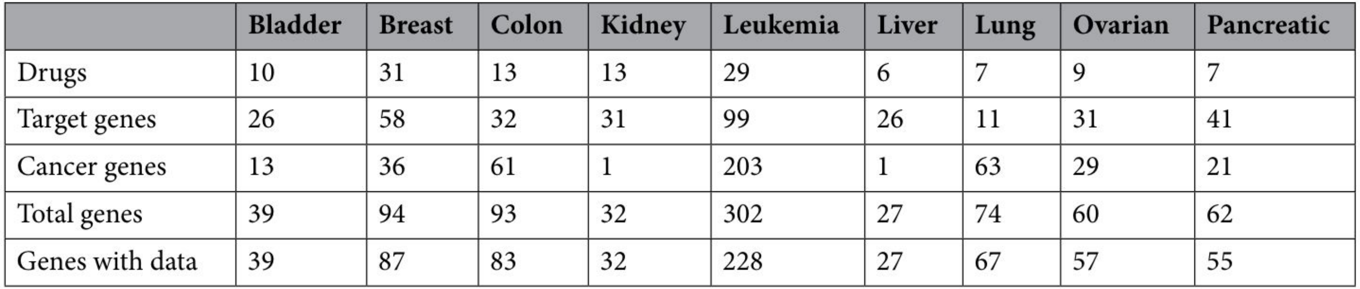
Data Collecting and Processing - Genes and Labels

POSITIVE

1:1

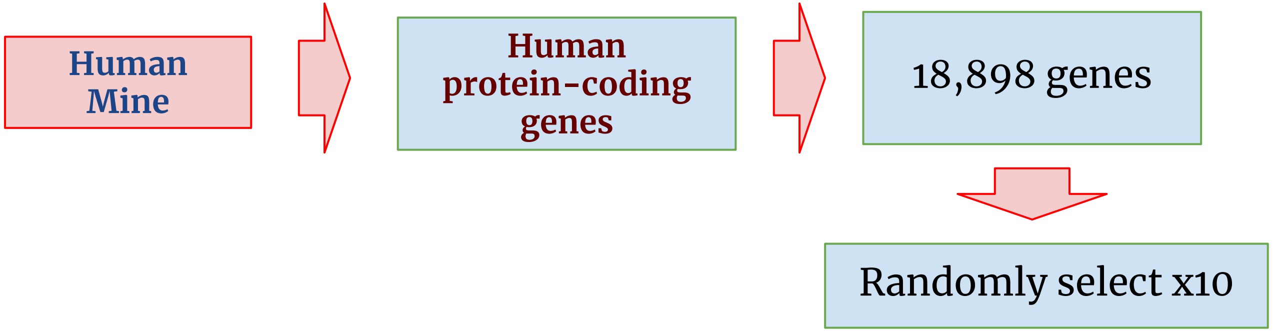
NEGATIVE

FDA-approved drugs for each cancer - US National Cancer Institute



	Bladder	Breast	Colon	Kidney	Leukemia	Liver	Lung	Ovarian	Pancreatic
Drugs	10	31	13	13	29	6	7	9	7
Target genes	26	58	32	31	99	26	11	31	41
Cancer genes	13	36	61	1	203	1	63	29	21
Total genes	39	94	93	32	302	27	74	60	62
Genes with data	39	87	83	32	228	27	67	57	55

Bazaga, A., Leggate, D., & Weisser, H. (2020).



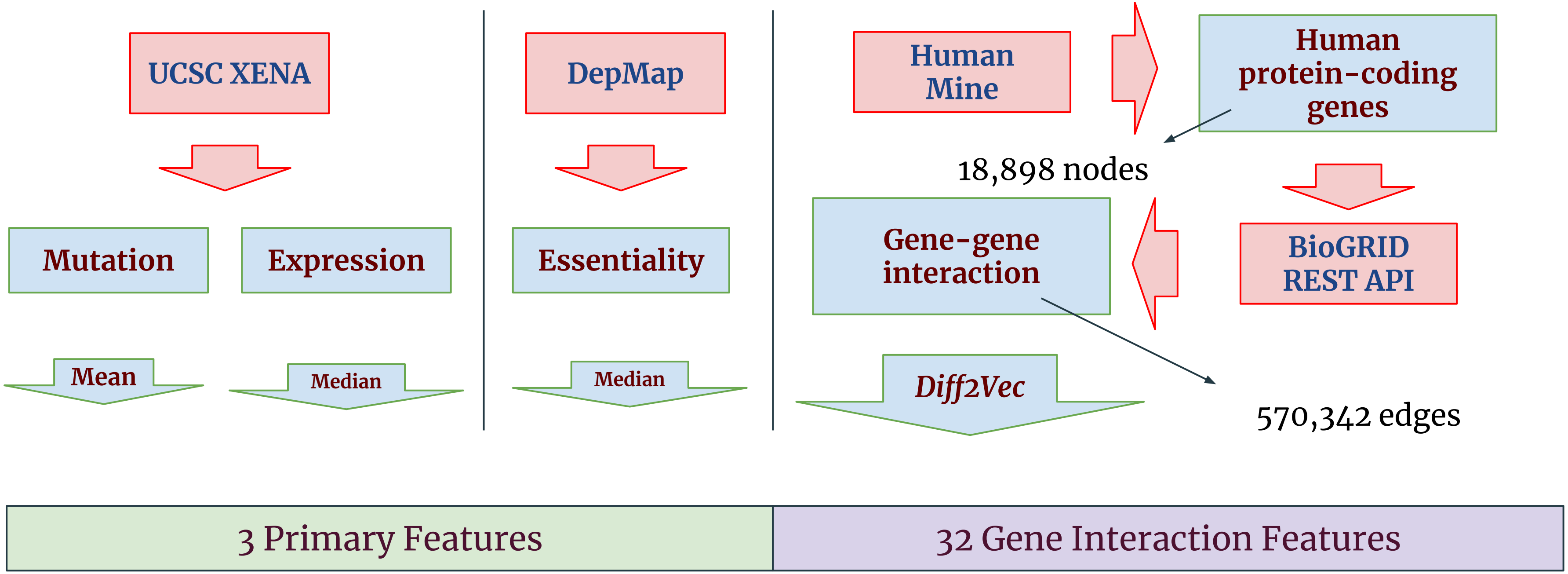
Human Mine

Human protein-coding genes

18,898 genes

Randomly select x10

Data Collecting and Processing - Features



Per cancer type

Regardless of cancer type

Multivariate Feature Selections

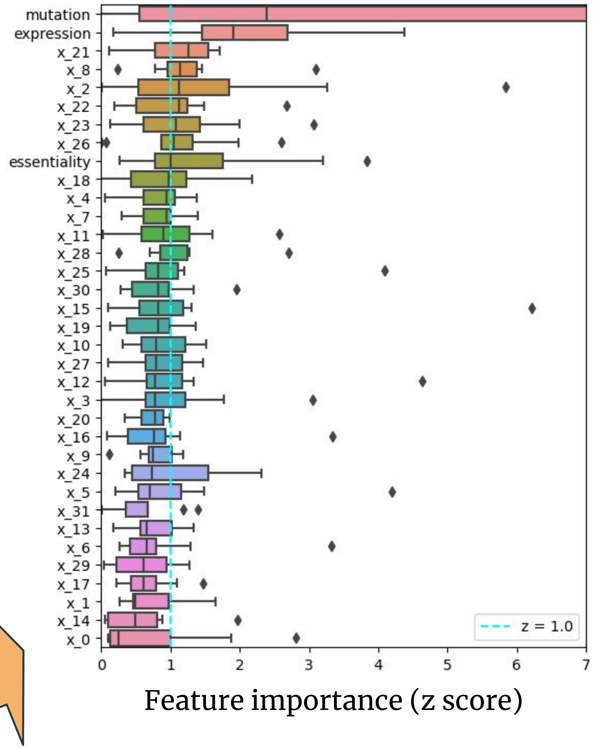
Importance using RF

Shuffle labels 100 times, then calculate importance using RF

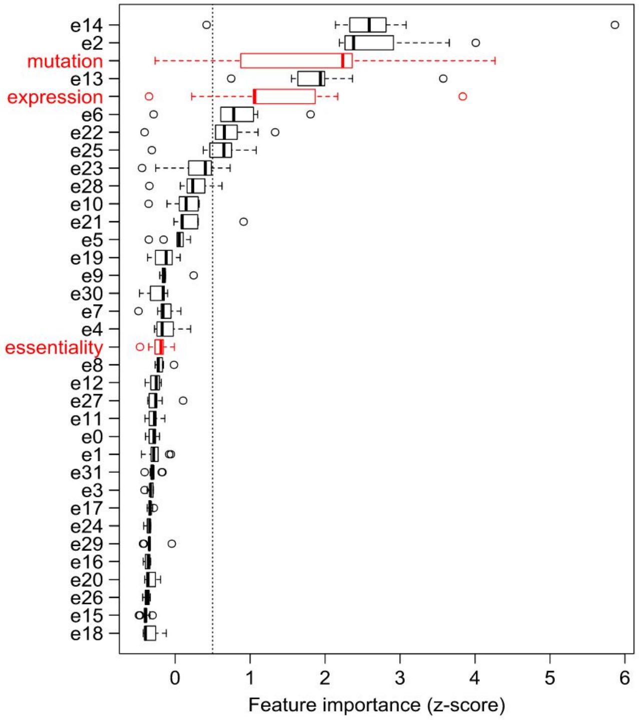
Normal distribution of feature importance

Z score of each feature importance of each cancer type

This work

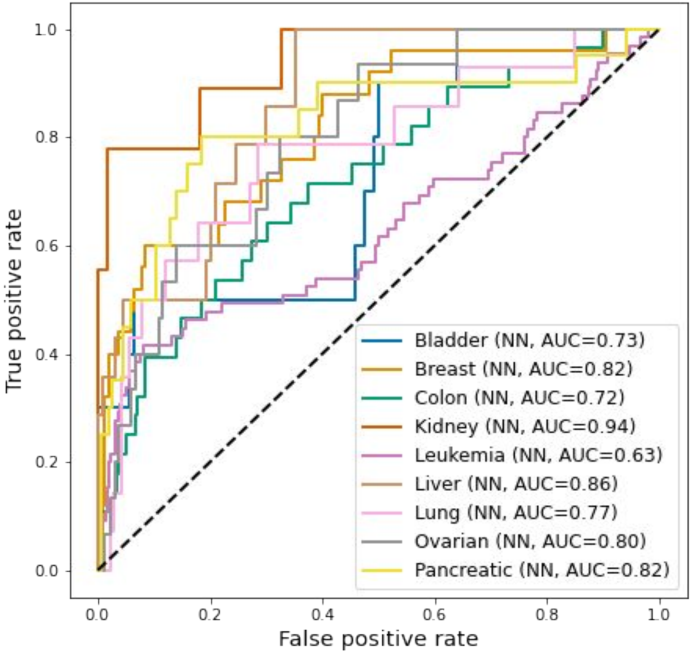


Literature

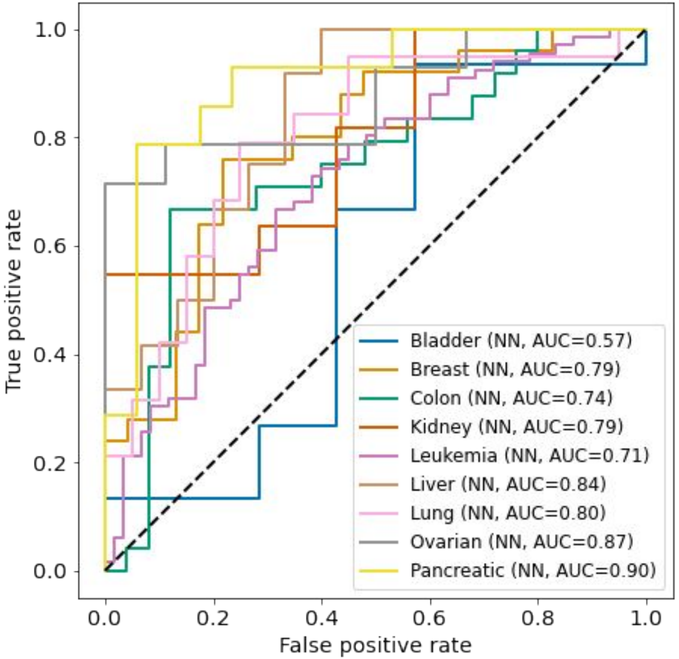


Model
Artificial Neural Network

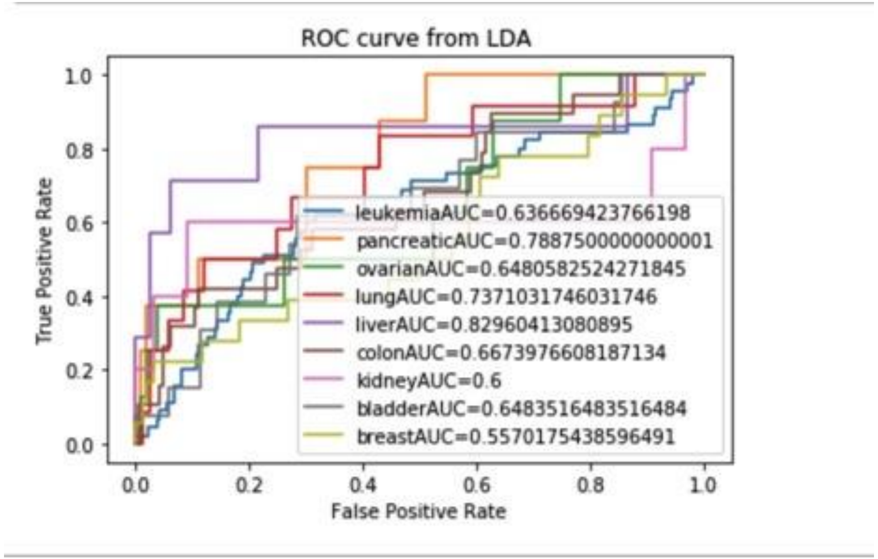
32 features

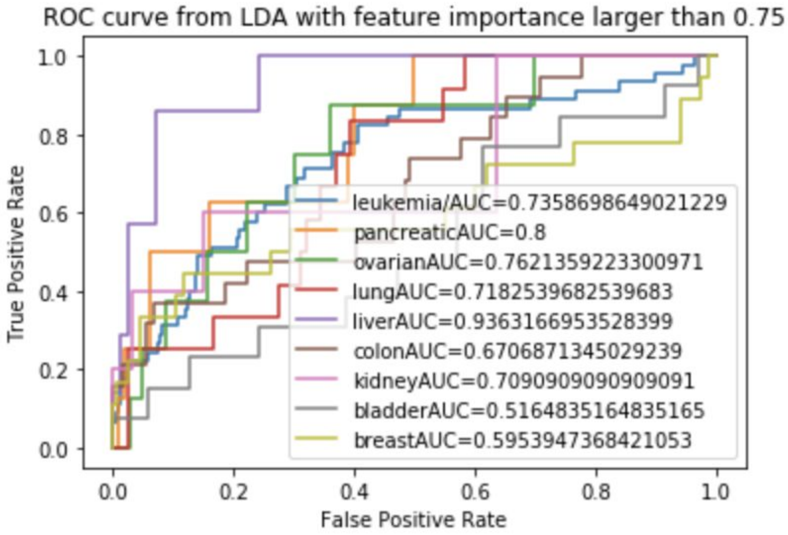


35 features (with primary)



Model 2
Linear Discriminant Analysis(LDA)

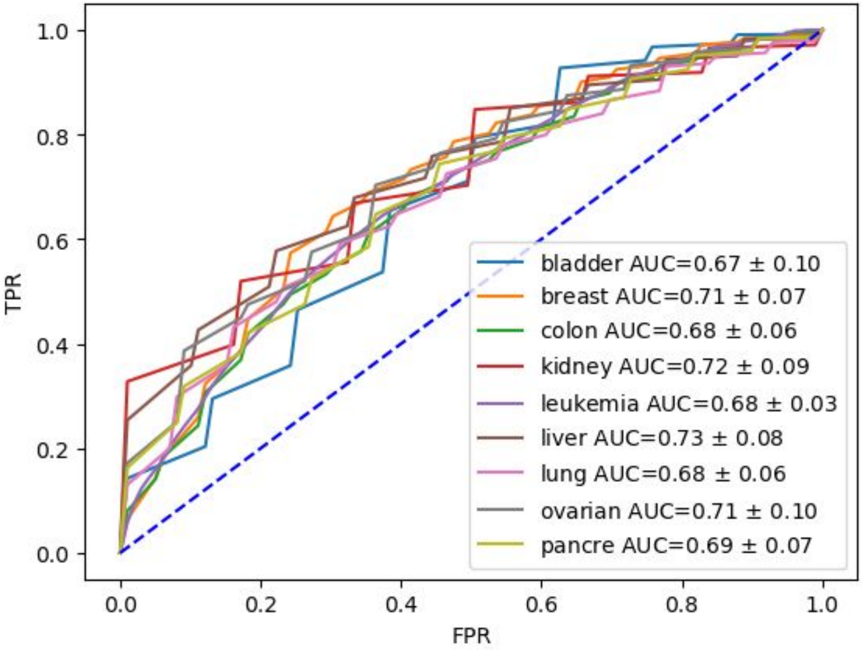




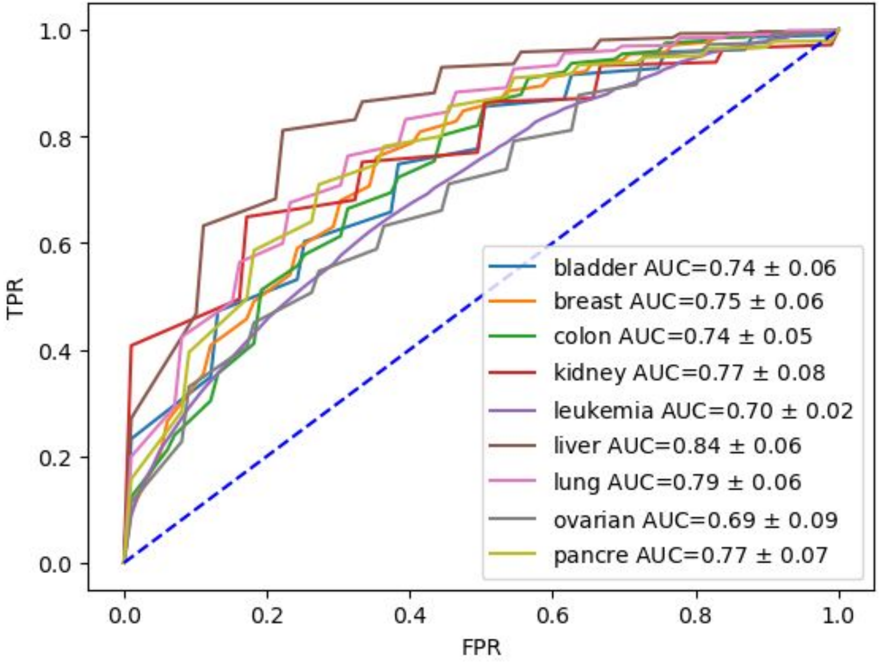
WITH 3 primary features

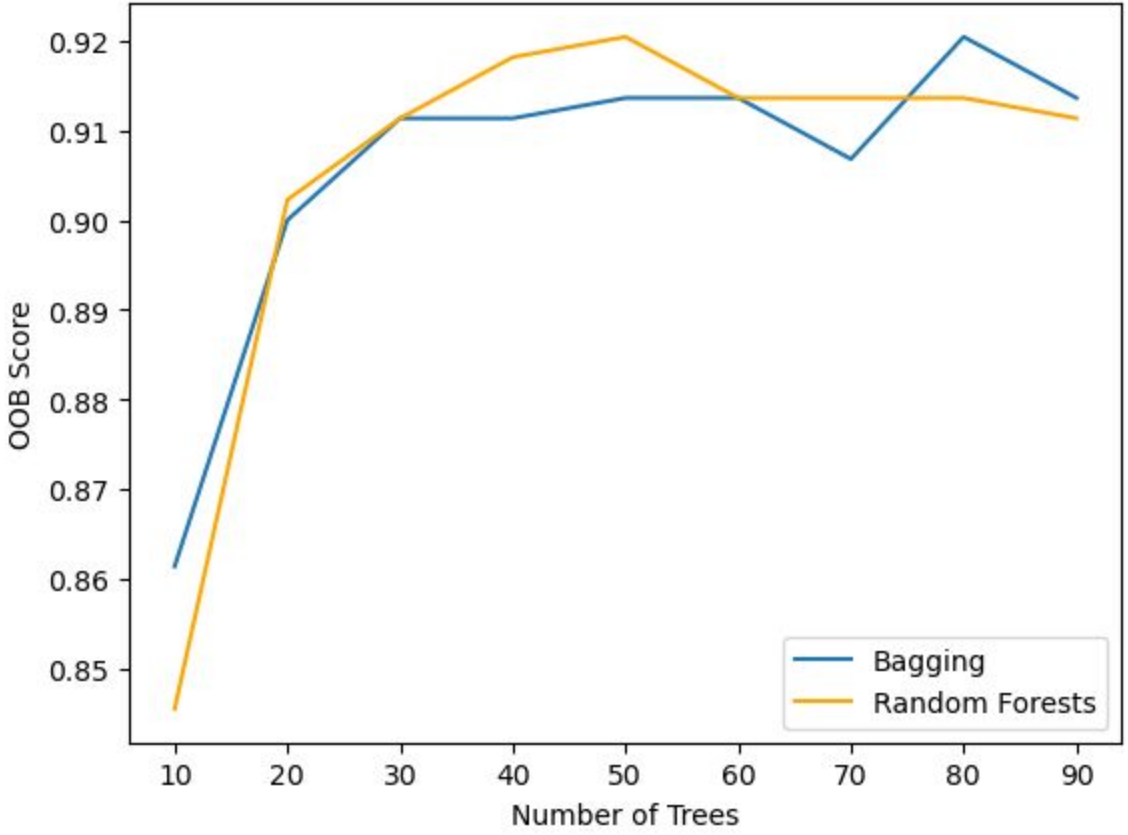
Model 3
Random Forest

32 features



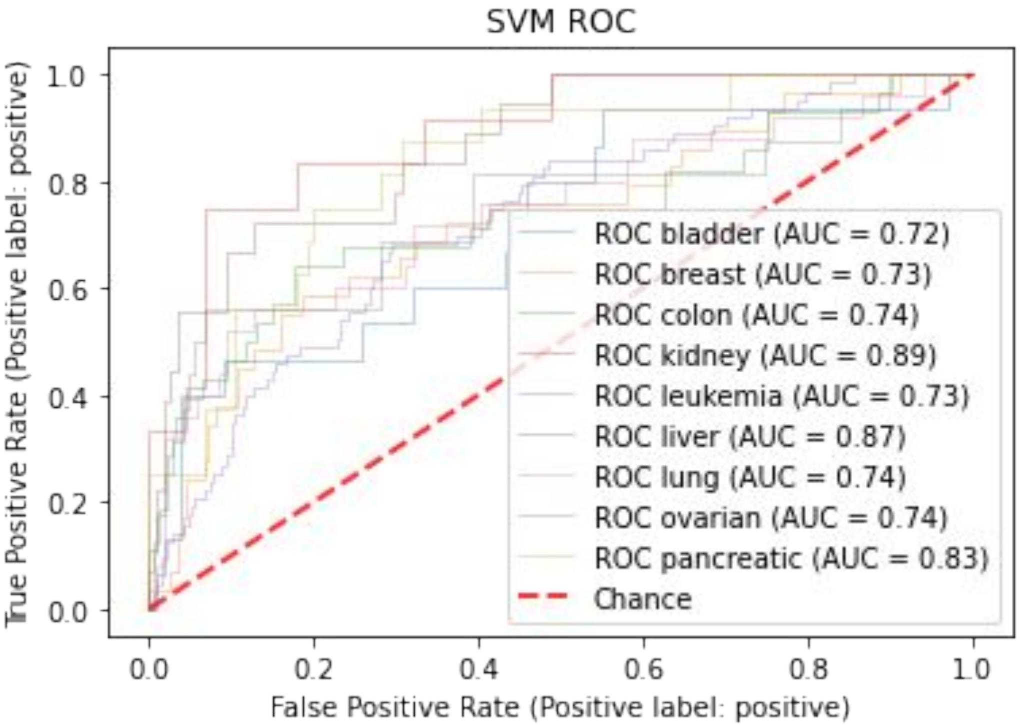
35 features (with primary)



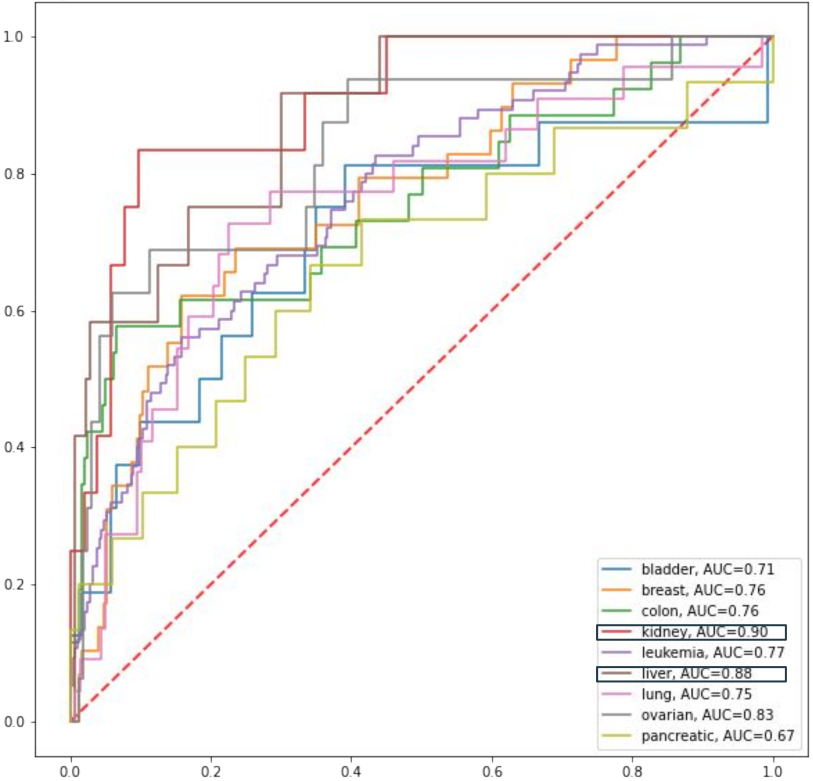

0.92
0.91
0.90
0.89
0.88
0.87
0.86
0.85
OOB Score
10
20
30
40
50
60
70
80
90
Number of Trees
Bagging
Random Forests

SVM

With 35 features including 3 primary features



Model 6
Logistic Regression



Logistic Regression method
Receiver operating characteristic (ROC) curve

Discussion

- Need different models for different lineages
- Compared to the paper’s drug based data set, the newer gene dependency data set has both different feature importance and model accuracy.
- Addition of network based features can add bias towards genes that are well studied.

Future Work

- Apply different labelling methodology using CRISPR KO.
- Find new target genes using CRISPR KO Labelling and compare results with drug-based labelling.
- Incorporating other features can help reduce potential bias.
- A semi-supervised model can help incorporate genes with “unknown” status into our model.

Questions?