These aren’t the SNPs you’re looking for: Jedi deepMIND tricks for eye color prediction

STR Crazy
James Talwar and Adam Klie

Phenotype prediction with SNPs

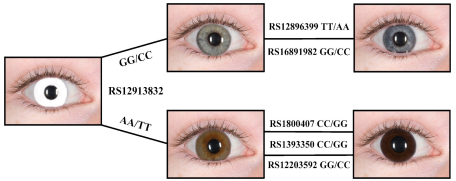
Current approaches often rely on linear, additive models of SNPs based on GWAS statistics

- Capture only a subset of the heritable variation
- Miss interplay between SNPs (epistasis), low effect size SNPs

$$p_{blue} = \frac{e^{\alpha_1 + \sum_k \beta_{1,k} X_k}}{1 + e^{\alpha_1 + \sum_k \beta_{1,k} X_k} + e^{\alpha_2 + \sum_k \beta_{2,k} X_k}}$$

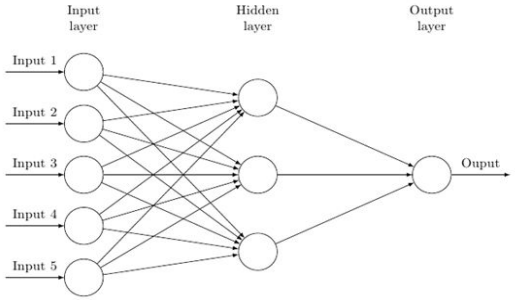
$$p_{other} = \frac{e^{\alpha_2 + \sum_k \beta_{2,k} X_k}}{1 + e^{\alpha_1 + \sum_k \beta_{1,k} X_k} + e^{\alpha_2 + \sum_k \beta_{7,k} X_k}}$$

$$p_{brown} = 1 - p_{blue} - p_{other}$$



Deep neural networks have potential to overcome these limitations

- Potential to handle larger number of SNPs
- Can capture complex, non-linear interactions

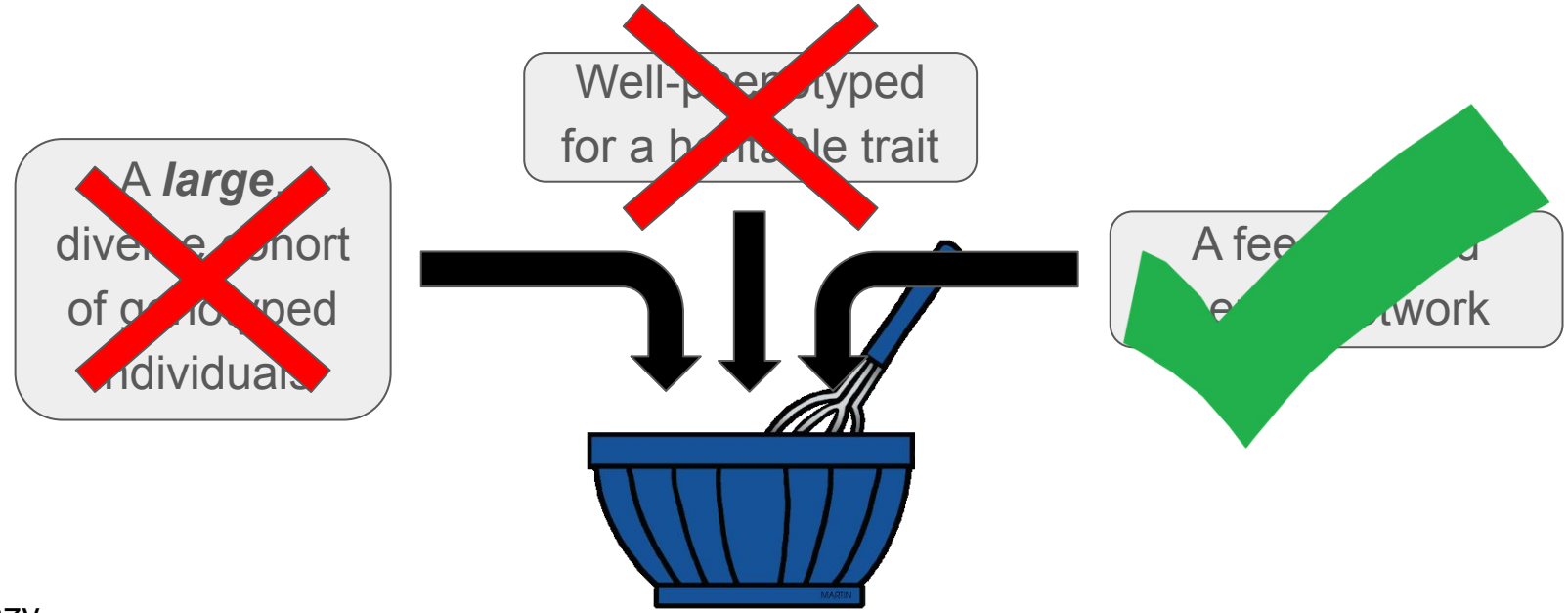


Walsh, Susan, et al. "IrisPlex: A Sensitive DNA Tool for Accurate Prediction of Blue and Brown Eye Colour in the Absence of Ancestry Information."

STR Crazy

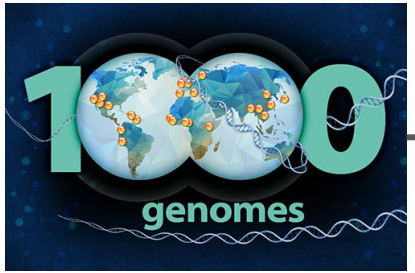
Motivation Methods Results Summary

Goal: Train a feedforward network to predict a phenotype with high accuracy



STR Crazy

1000Genomes



2504 diverse individuals
No phenotypes

IRISPLEX

2504 Samples	
Brown	2230
Blue	268
Other	6

COMMON SNPS

DTC genetic test database

590 samples	
Brown	312
Blue	194
Green	84

open SNP

5593 individuals
622 phenotypes
Heterogeneous dataset

276,563 SNPs

★

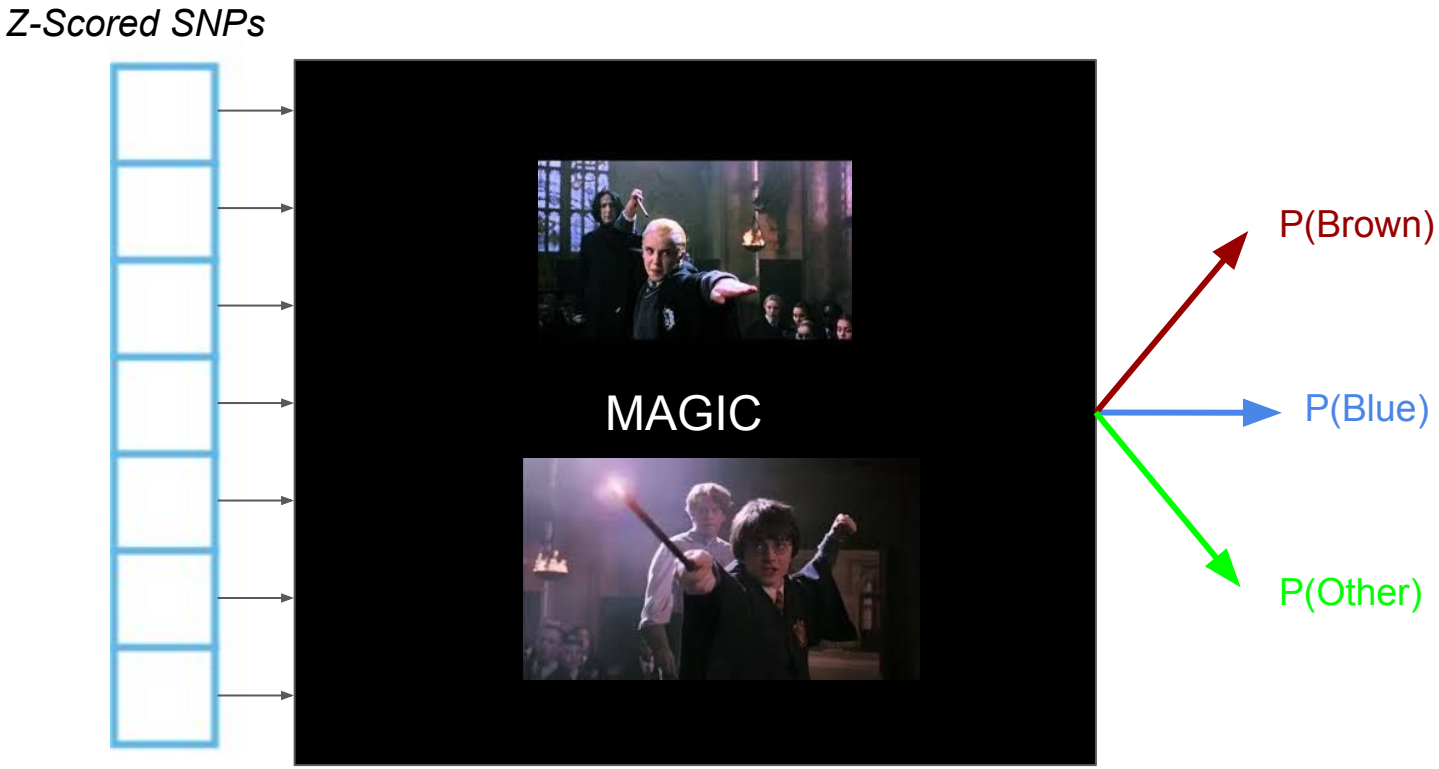
1000 SNPs by p-value

TRAINING (80%)		VAL (20%)	
Brown	1784	Brown	446
Blue	214	Blue	54
Other	5	Other	1

TEST	
Brown	312
Blue	194
Other	84

Motivation Methods Results Summary

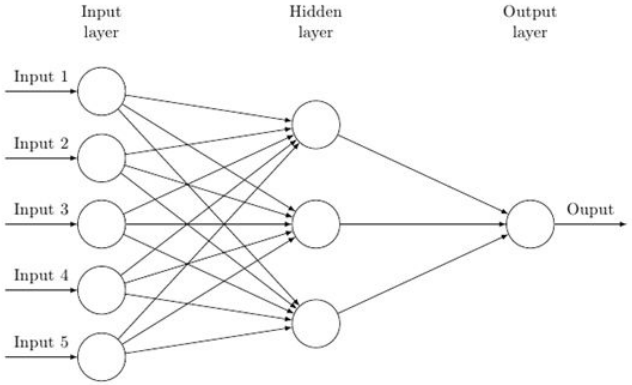
Neural Network Architecture:

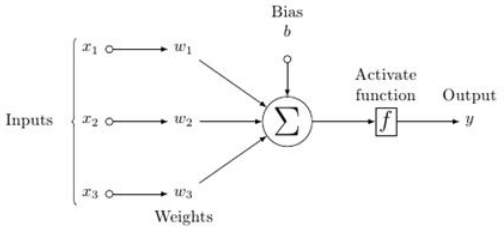


STR Crazy

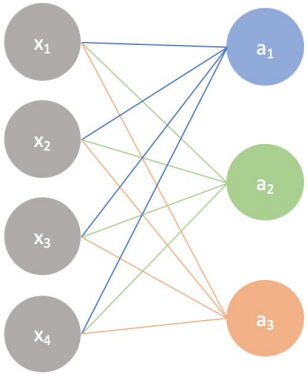
Motivation Methods Results Summary

How the *Magic* Happens:





Input Layer **Hidden Layer**



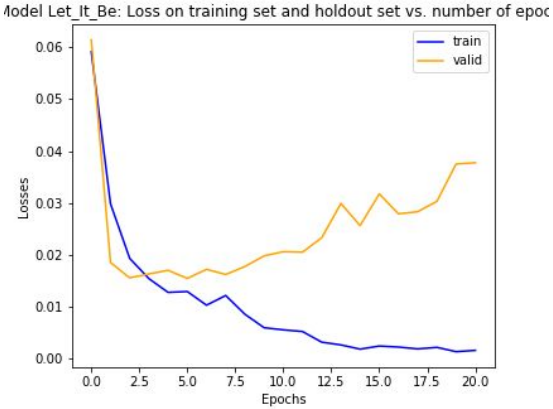
$$\begin{bmatrix} w_1 & w_2 & w_3 & w_4 \\ w_1 & w_2 & w_3 & w_4 \\ w_1 & w_2 & w_3 & w_4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} b \\ b \\ b \end{bmatrix} = \begin{bmatrix} w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + b \\ w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + b \\ w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + b \end{bmatrix} \xrightarrow{activation} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

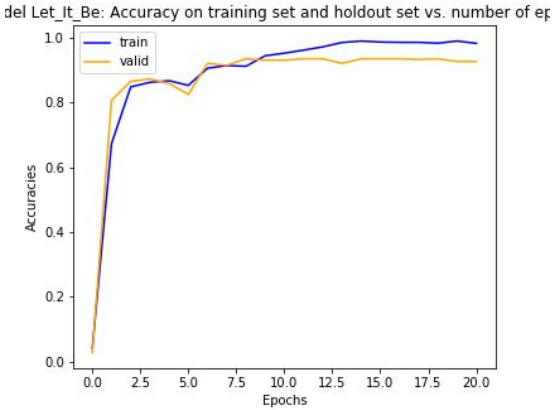
STR Crazy

Motivation Methods Results Summary

What's in the Box?: Hyperparameter Searching - All SNPs

Model	# Hidden	Layer Widths	Train Accuracy	Validation Accuracy
Let It Be	3	512; 128; 32	98.5%	93.4%
Double Trouble	2	1024; 512	97.9%	93.4%
Dos Equis	2	512; 256	98.7%	93.4%
Winter Is Coming	3	512; 256; 256	98.6%	93.4%
Cinco De Mayo	5	512; 256; 128; 64; 32	97.8%	93.4%



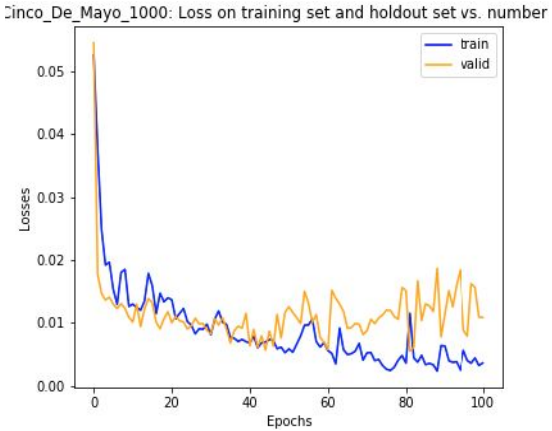


STR Crazy

Motivation Methods Results Summary

What's in the Box?: Hyperparameter Searching - 1000 SNPs

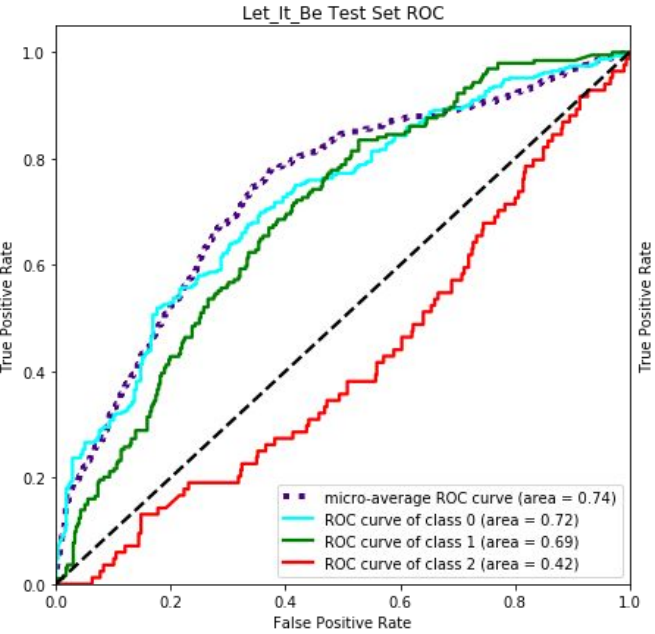
Model	# Hidden	Layer Widths	Train Accuracy	Validation Accuracy
Cinco De Mayo 1000	5	512; 256;128; 64; 32	96.4%	96.4%
Trisomy 1000	3	512; 256; 128	98.2%	95.8%
George 1000	1	128	99.5%	95.8%
Paul 1000	1	512	99.8%	95.8%
Glaucoma 1000	5	512; 512; 512; 512; 512	97.8%	95.6%

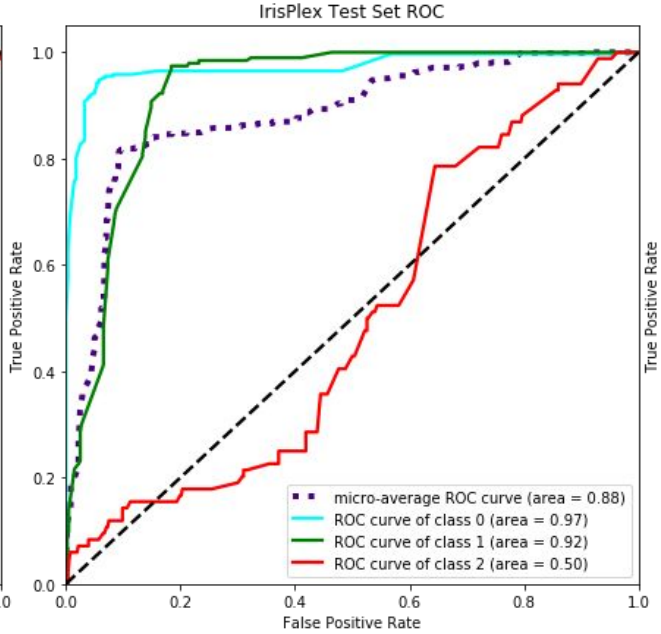


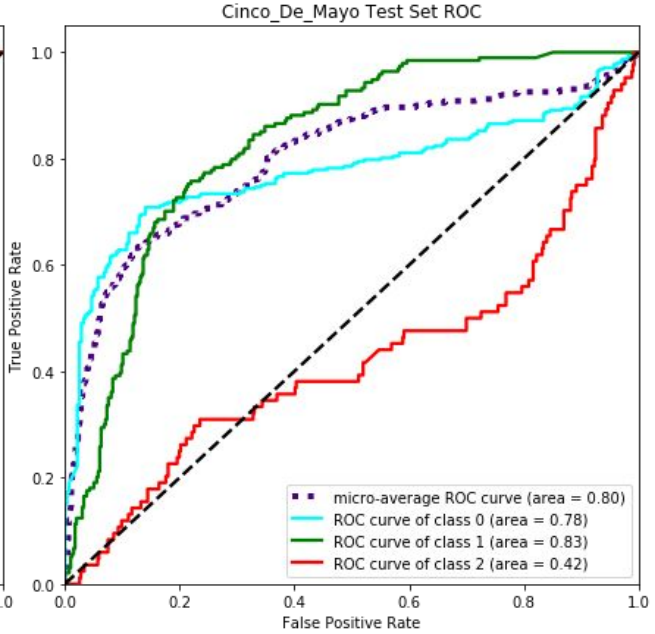
STR Crazy

Motivation Methods Results Summary

Make It (T)rain: Top Models vs. Irisplex on OpenSNP







All SNPs Test Accuracy: 55.1%

Brown 54.2%; **Blue** 80.4%; **Other** 0%

IrisPlex Test Accuracy: 81.2%

Brown 90.7%; **Blue** 98.5%; **Other** 6%

1000 SNPs Test Accuracy: 68.6%

Brown 76.6%; **Blue** 85.6%; **Other** 0%

STR Crazy

Motivation Methods Results Summary

The Test Set STRikes Back: Limitations and Challenges

- No ground truth training labels: Training was based on IrisPlex labels for most probable class as opposed to true labels
- Test set labels were self-reported and messy
- Test set Inconsistencies in SNPs across genotyping arrays:
  - Tradeoffs between filtering and number of SNPs available for training
  - Missing SNP problem: If no SNP reported we ignored contributions of that SNP in the prediction for that individual

STR Crazy

The Return of the Jedi: Summary and Future Directions

Summary:

- We constructed a flexible neural network architecture for multi-class prediction that can be generalized and adapted to any task

Future Directions:

- Gauge performance of our proof of conSNPt idea on a large phenotyped dataset
- Broader SNP selection (i.e., Use p-value in hyperparameter search)
- Adapt model to handle noise in inputs to prevent poor performance when certain SNPs aren't reported for an individual

STR Crazy

Motivation Methods Results Summary

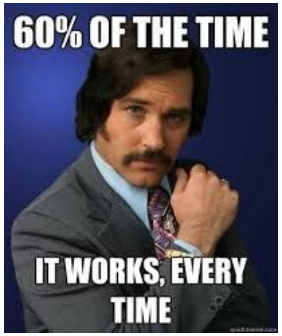
Acknowledgments

Dr. Gymrek

Shubham

STR Crazy

Questions?



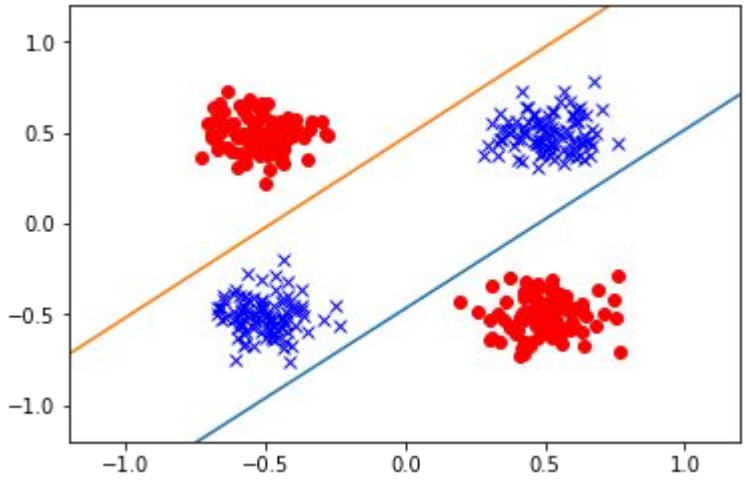
"60% of the time, this model works every time"
- James Talwar

"I never want to see a panda again"
- Adam Klie

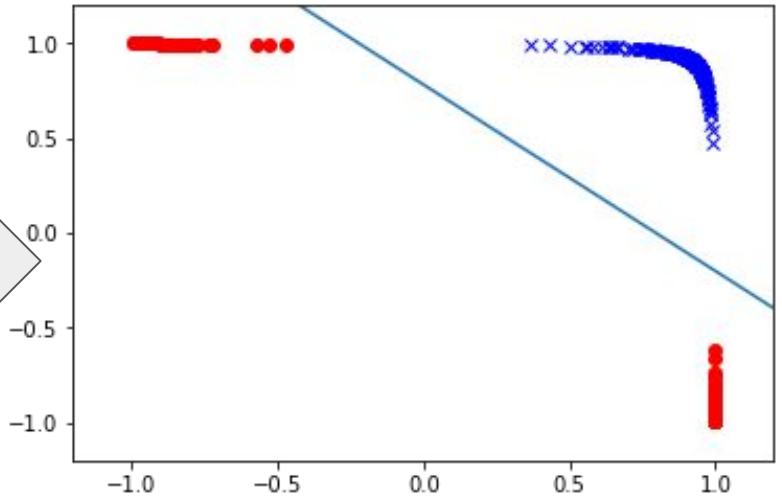
Supplementary Slides

STR Crazy

Non-linear Transformation

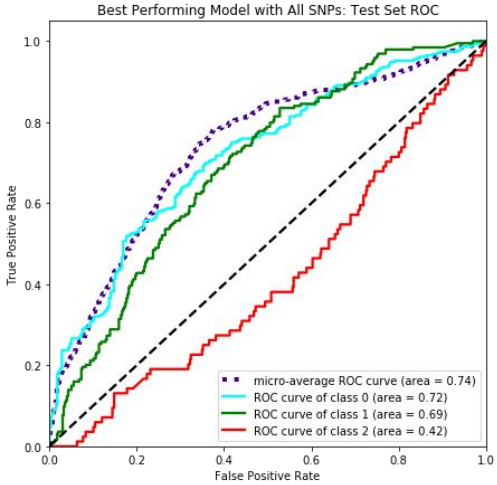


Non-Linear Activation



STR Crazy

Curves or metrics comparing accuracy on IrisPlex vs the accuracy of our model



STR Crazy