Mechanisms of Action (MoA) Prediction

Can you improve the algorithm that classifies drugs based on their biological activity?

JIE LIU, PH.D.

The Project

- A Kaggle competition organized by MIT and Harvard
- The dataset combines gene expression and cell viability data as measurements of human celllular responses to drug treatment
- The task is to use this dataset to develop a machine learning model that automatically labels any new drug as one or more MoA types

My Objective

- build all kinds of classification models and compare their performances

The Challenges

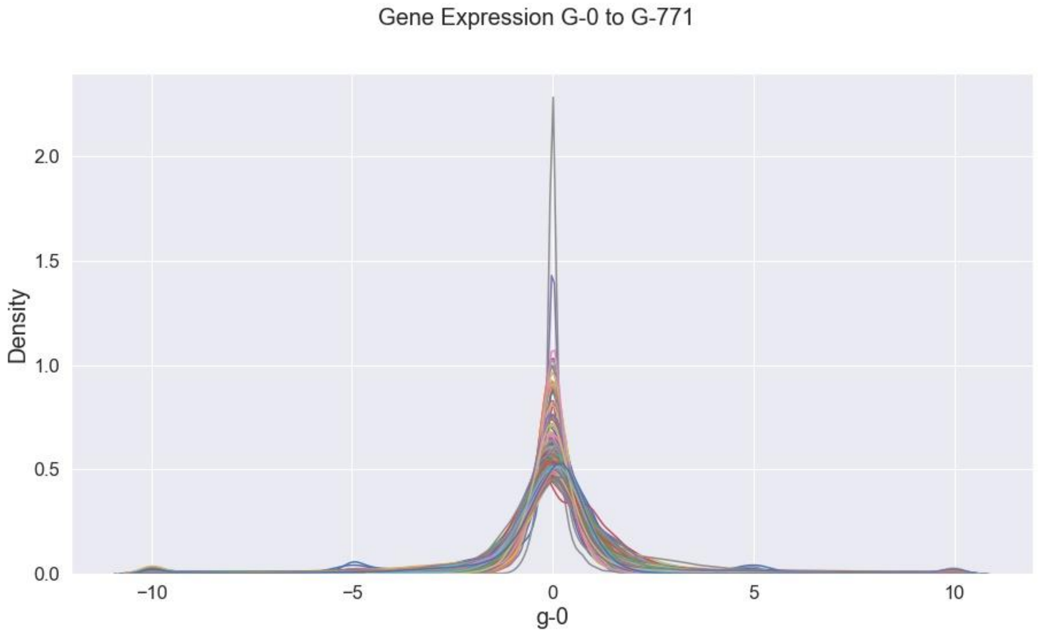
- A multi-label classification, 206 labels in total
- Custom evaluation metric required, logarithmic loss function, formula provided

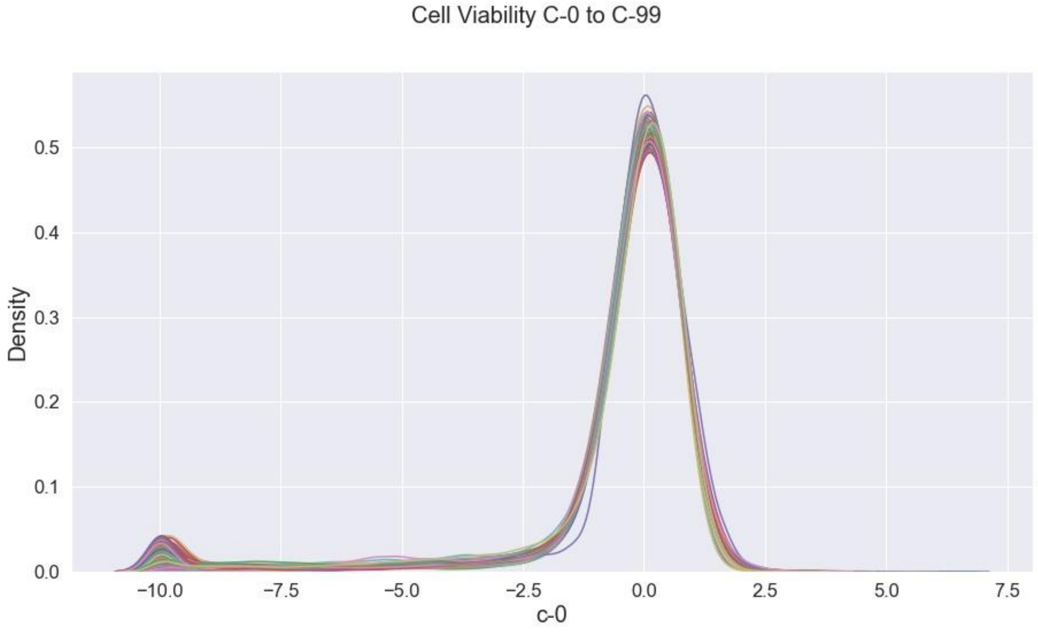
$$\text{score} = -\frac{1}{M}\sum_{m=1}^{M}\frac{1}{N}\sum_{i=1}^{N}\left[y_{i,m}\log(\hat{y}_{i,m}) + (1 - y_{i,m})\log(1 - \hat{y}_{i,m})\right]$$

- Features are highly correlated to each other

Exploratory Data Analysis

- 17860 drug samples, 875 features including gene expression and cell viability patterns in response to drug treatment, 206 labels
- 872 numerical features
- Data pre-normalized, following normal-like distributions



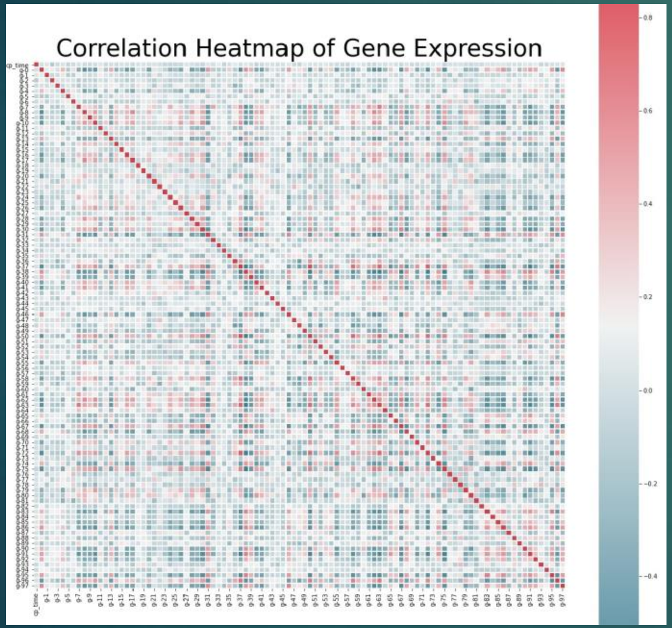


Exploratory Data Analysis (cont.)

High correlation suggests PCA dimensionality reduction

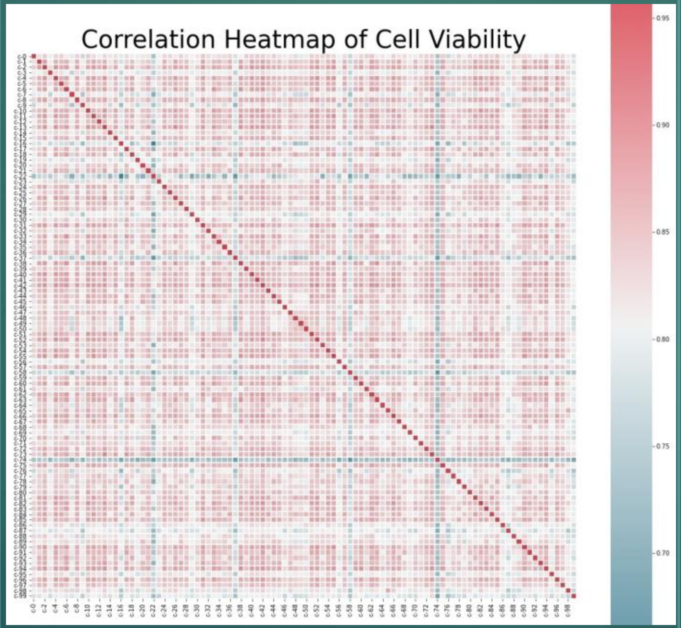
Gene Expression

Moderately Correlated



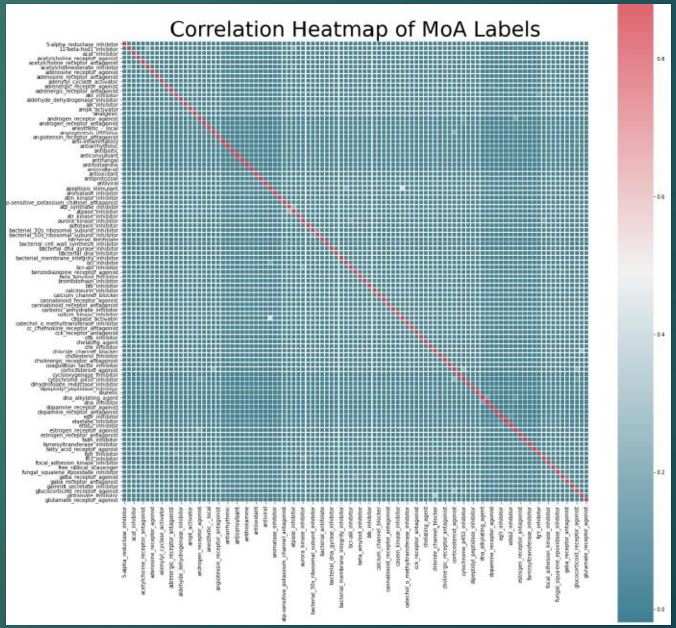
Cell Viability

Highlly Correlated



MoA Labels

Independent to each other



Data Processing

- No need to normalize data again
- Data split into 60% train, 20% validation and 20% holdout

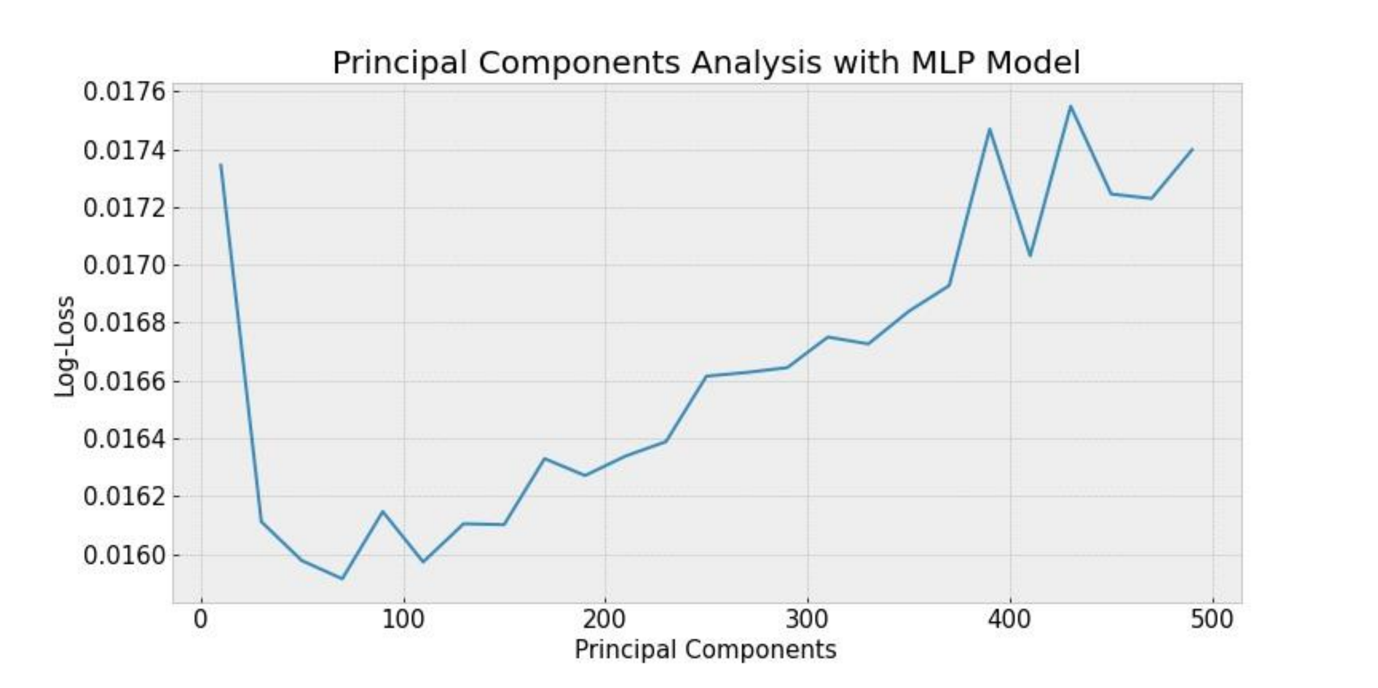
First Two Neural Network Models

Baseline Model	Neural Network Structure	Parameters	Log-loss
Multilayer Perceptron	1 input, 1 hidden, 1 output layers	91,470	0.0183
1-D Convolutional NN	1 input, 2 hidden, 1 output layers	905,102	0.0194

- MLP model yields smaller thus better log-loss.

Principal Component Analysis (PCA)

Dimensionality Reduction



Original feature number: 875

n_components scan:10 to 500

Best log-loss: at 70

Log-loss MLP: **0.0159** (down from **0.0183**)

Log-loss 1D-CNN: **0.0180** (down from **0.0194**)

15 Classification Models

MLP and All kinds of Convolutional Neural Networks

	MLP	1D-CNN	AlexNet	LeNet-5	VGG-16 Net	ResNet	Inception Net
Parameters	91,470	905,102	23,337,214	116,110	28,828,174	340,430	5,235,936
Log-loss	0.0159	0.0180	0.0199	0.0170	0.0176	0.0191	0.0189
	*Best			*			

Other Classification Models

	Random Forest	SVC	KNN	XGBoost	Adaboost	Logistic Regression	GaussianNB	Decision Tree
Log-loss	0.0165	0.0167	0.0168	0.0170	0.0172	0.0177	0.0434	0.0684
	*	*	*	*				

* Six models selected for further optimization

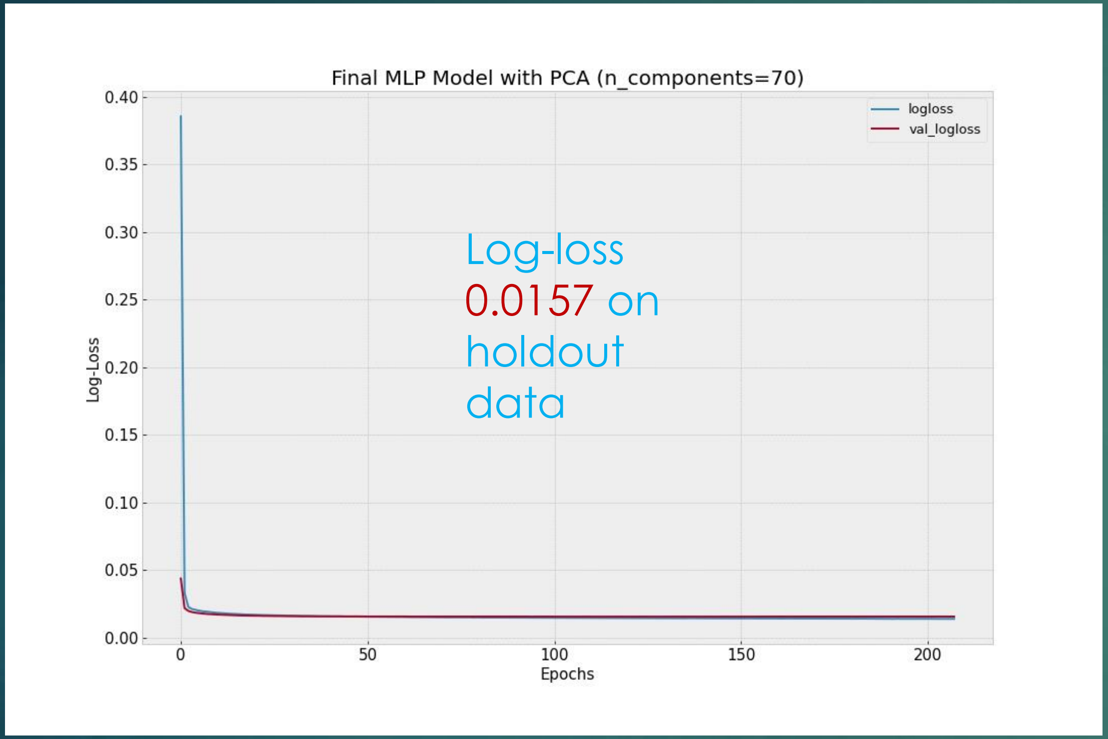
Hyperparameter Tuning - Hyperopt

- A powerful python library for hyperparameter tuning, using Bayesian optimization algorithm.
- Allowing the optimization of hundreds of parameters efficiently.
- Highly recommended over Sklearn RandomizedSearch and GridSearch optimization.

	Log-loss Before Optimization	**Log-loss After Optimization**
Multilayer Perceptron (MLP)	0.0159	0.0156
LeNet-5	0.0170	0.0161
C-Support Vector Classification (SVC)	0.0167	0.0164
Gradient Boost XGBoost	0.0170	0.0166
RandomForest	0.0165	0.0166
K Nearest Neighbours (KNN)	0.0168	0.0168

My Final MLP Model

1 Input layer, 2 Hidden Layers, 1 Output Layer, Batch Normalization and Dropout



#	Δpub	Team Name	Notebook	Team Members	Score	Entries	Last
1	▲6	Hungry for gold	Fork of Blending w...		0.01599	6	5mo
2	▲2	tmp			0.01599	6	5mo
3	▲3	Duck Quake			0.01600	116	5mo
4	▲8	Kanna Hashimoto with Fri...	nn-svm-tabnet-...		0.01600	50	5mo
5	▲9	YuyaAnna			0.01602	130	5mo
6	▲10	MooooooooA			0.01602	37	5mo
7	▲10	Cakey			0.01603	243	5mo
8	▲3	Caio Camilli			0.01603	202	5mo
9	▲11	The Slippery Appraisals			0.01603	261	5mo
10	▲3	Thomas Yokota			0.01603	46	5mo

Top 10 Winners of the Competition

Conclusions

- The simple MLP neural network model works best for my project.

Techniques Most Helpful

- Principle component analysis (PCA)
- Hyperparameter optimization with Hyperopt

Jie Liu, Ph.D.

- jliu1999@gmail.com
- Tel: 917-306-8708
- github.com/jliu1999
- linkedin.com/in/jieliu1999