Predicting drug sensitivity using (epi)genomic marks

Elysia Chou, Keefer Chern, Alexander Chang

May 1, 2019

Introduction

Cancer cell lines are widely used to study drug efficacy in vitro.

Aims:

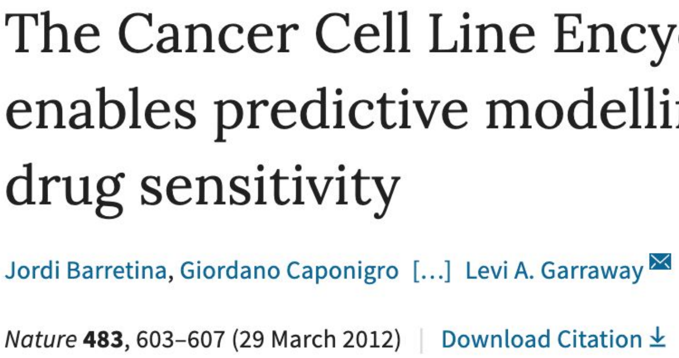
Predictively model drug efficacy using elastic net regression with either

1. Epigenomic features (histone modifications)

   or

2. Genomic features (gene expression)

Motivation: compare to prior work



The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity

Jordi Barretina, Giordano Caponigro [...] Levi A. Garraway ✉

Nature **483**, 603–607 (29 March 2012) | Download Citation ⤓

Paper's approach:

- Trained on all of the data
- Uses genomic data (not all public)

Our approach:

- Train on 80% of the data
- Explore relationship between (epi-) genomic profiles and drug sensitivity

Data

Features

- Global Chromatin Profiling
- Gene Expression

Labels

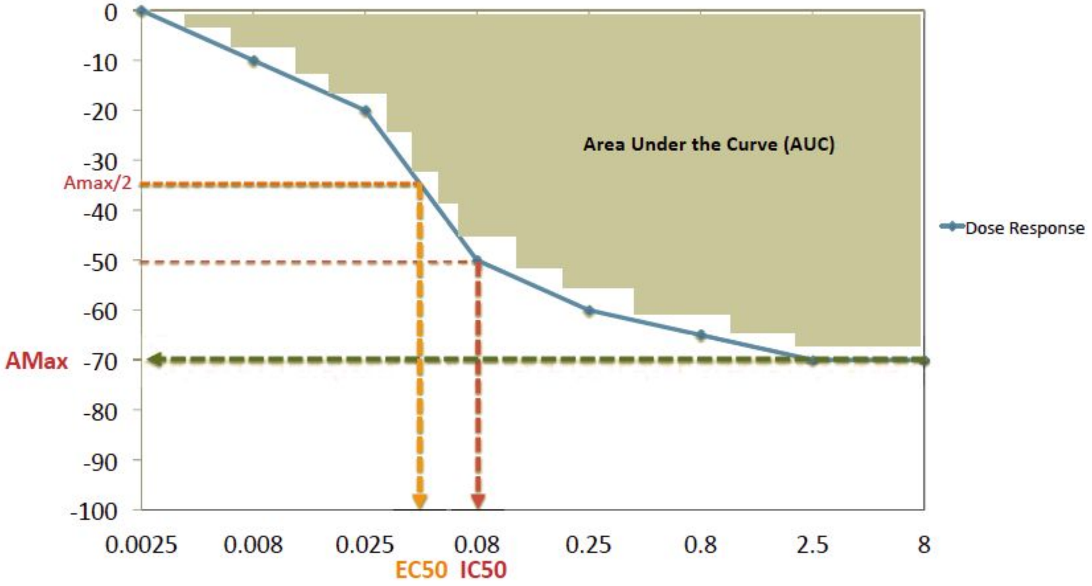
- Drug sensitivity for PD-0325901 (MEK inhibitor)

CCLE Cancer Cell Line Encyclopedia

Labels: Drug Sensitivity Data (PD-0325901)

Drug Response for a cancer cell line

Relative growth inhibition (%)

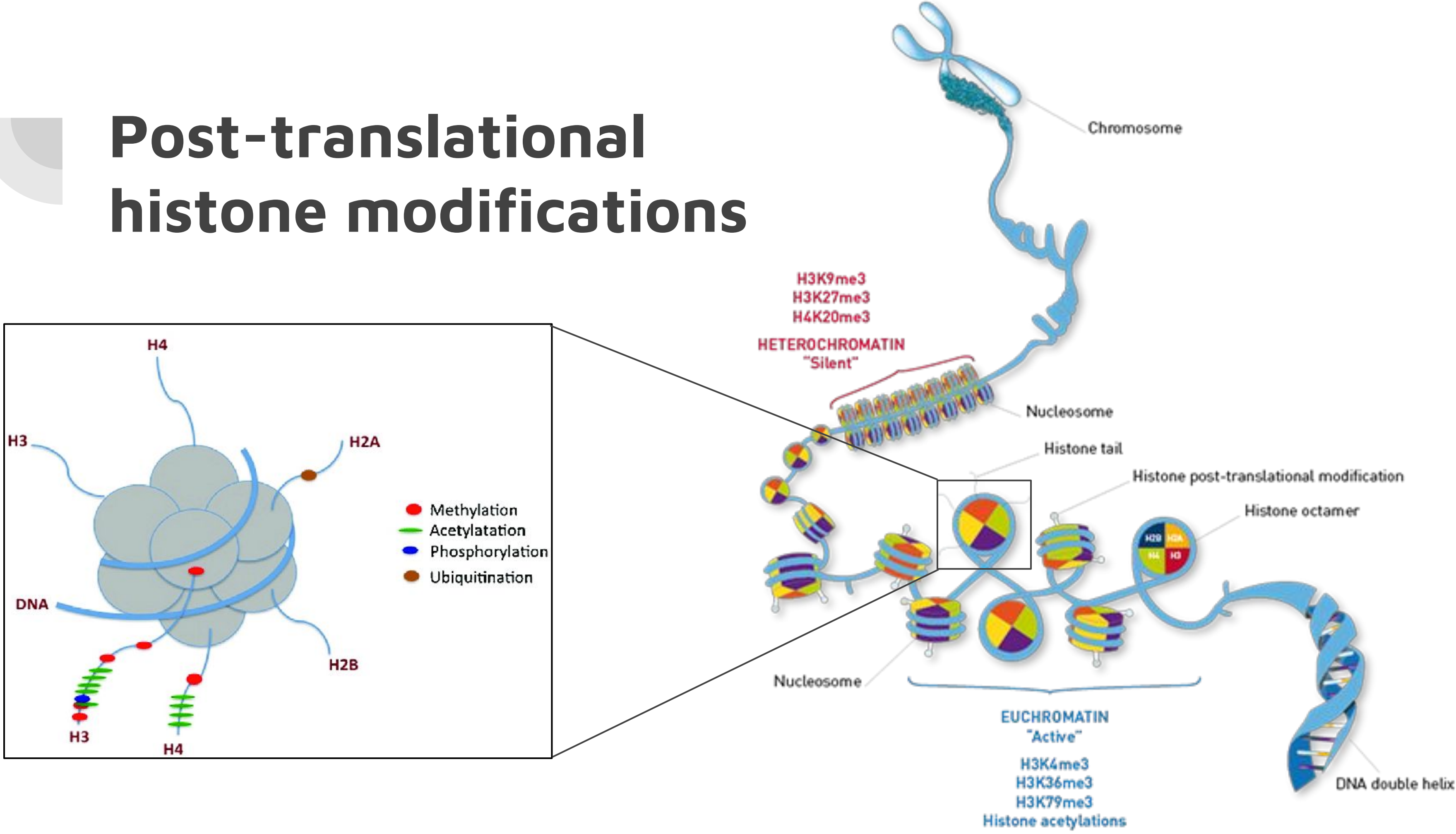


Drug concentration (in µM)

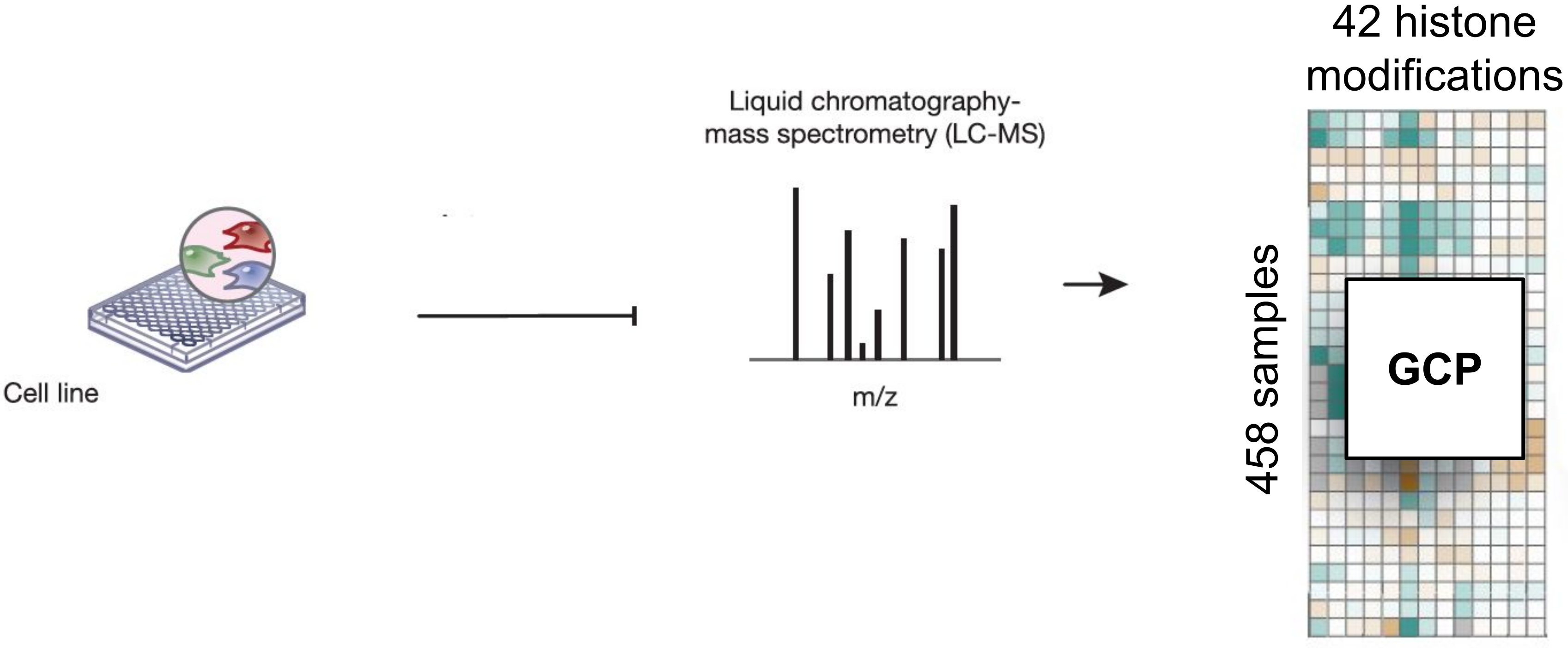
- EC50
- IC50
- Amax
- Activity Area/AUC

Global Chromatin Profiling Analysis

Post-translational histone modifications



Global Chromatin Profiling (GCP) Data



Cell line

Liquid chromatography-mass spectrometry (LC-MS)

m/z

42 histone modifications

458 samples

GCP

Resolve Missing Values

- If # NaNs ≤ 30: impute missing values with mean of the column
- Columns removed if # NaNs > 30
- Features remaining: 39
- Labels remaining: 3

Column	Column name	# of NaNs
2	H3K4me0	1
5	H3K4ac1	40
20	H3K18ac0K23ub1	162
34	H3K27ac1K36me0	3
35	H3K27ac1K36me1	15
36	H3K27ac1K36me2	5
37	H3K27ac1K36me3	1
40	H3K56me1	162
42	H3K79me1	1
43	H3K79me2	2

Method: Elastic Net Regression

- Linear regression with L1 and L2 regularization
- Supposed to be better at dealing with situations with correlations between parameters
- Loss function:

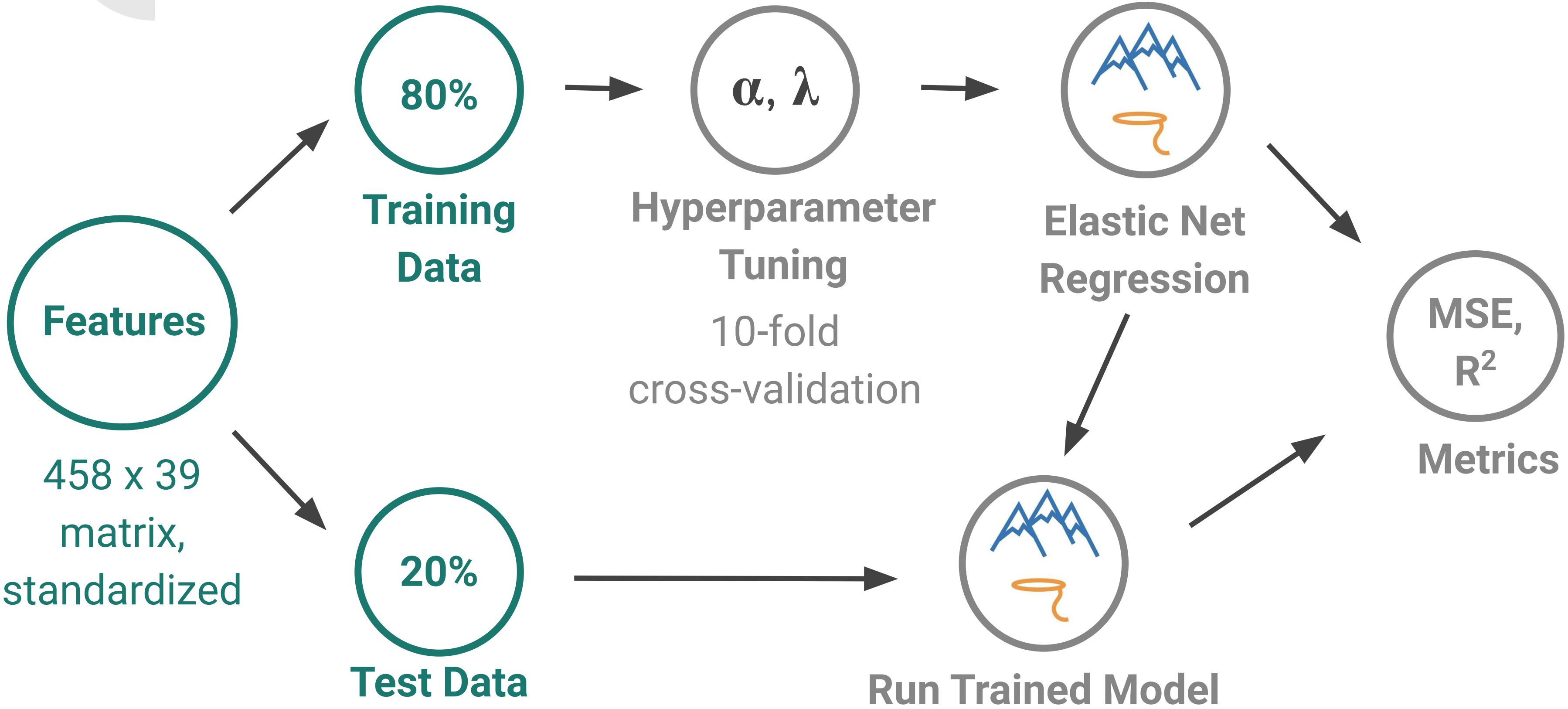
$$L_{enet}(\hat{\beta}) = \frac{\sum_{i-1}^{n} (y_i - x_i' \hat{\beta})^2}{2n} + \lambda(\frac{1-\alpha}{2} \sum_{j-1}^{m} \hat{\beta}_j^2 + \alpha \sum_{j-1}^{m} |\hat{\beta}_j|),$$

- **α** = mixing parameter (lower → less L1)
- **λ** = regularization parameter (0 = no regularization)

Hyperparameter Tuning for Elastic Net Regression

- 80/20 train/test split
- Determine optimal α and λ for our model using 10-fold cross-validation on training set
- MSE and R^2

Workflow



Features

458 x 39 matrix, standardized

80%

Training Data

20%

Test Data

α, λ

Hyperparameter Tuning

10-fold cross-validation

Elastic Net Regression

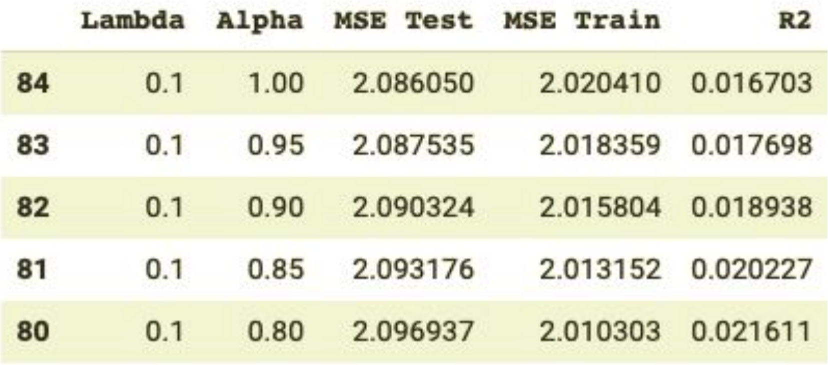
Run Trained Model

MSE, R^2

Metrics

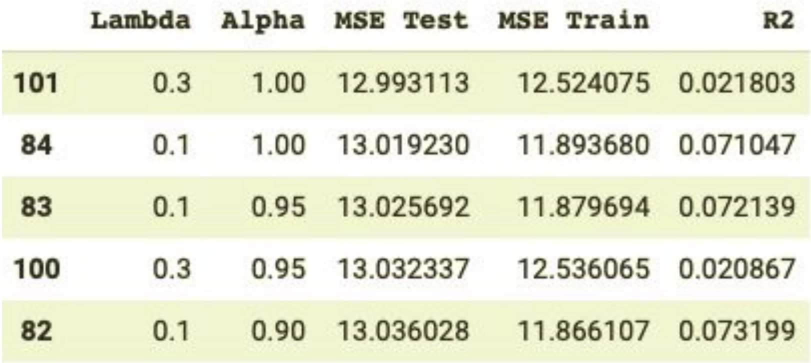
Results

From hyperparameter tuning



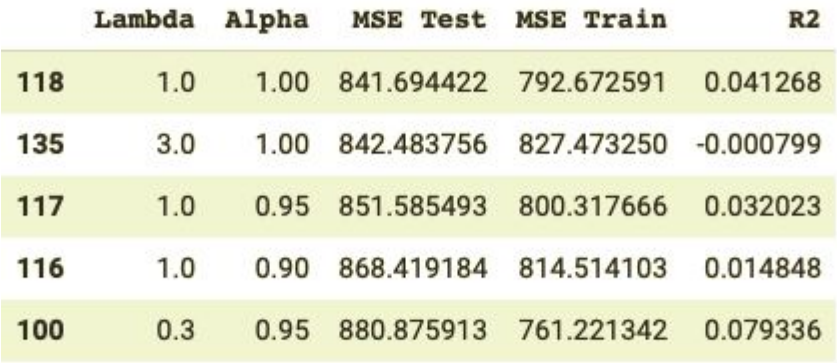
	Lambda	Alpha	MSE Test	MSE Train	R2
84	0.1	1.00	2.086050	2.020410	0.016703
83	0.1	0.95	2.087535	2.018359	0.017698
82	0.1	0.90	2.090324	2.015804	0.018938
81	0.1	0.85	2.093176	2.013152	0.020227
80	0.1	0.80	2.096937	2.010303	0.021611

Activity Area



	Lambda	Alpha	MSE Test	MSE Train	R2
101	0.3	1.00	12.993113	12.524075	0.021803
84	0.1	1.00	13.019230	11.893680	0.071047
83	0.1	0.95	13.025692	11.879694	0.072139
100	0.3	0.95	13.032337	12.536065	0.020867
82	0.1	0.90	13.036028	11.866107	0.073199

IC 50

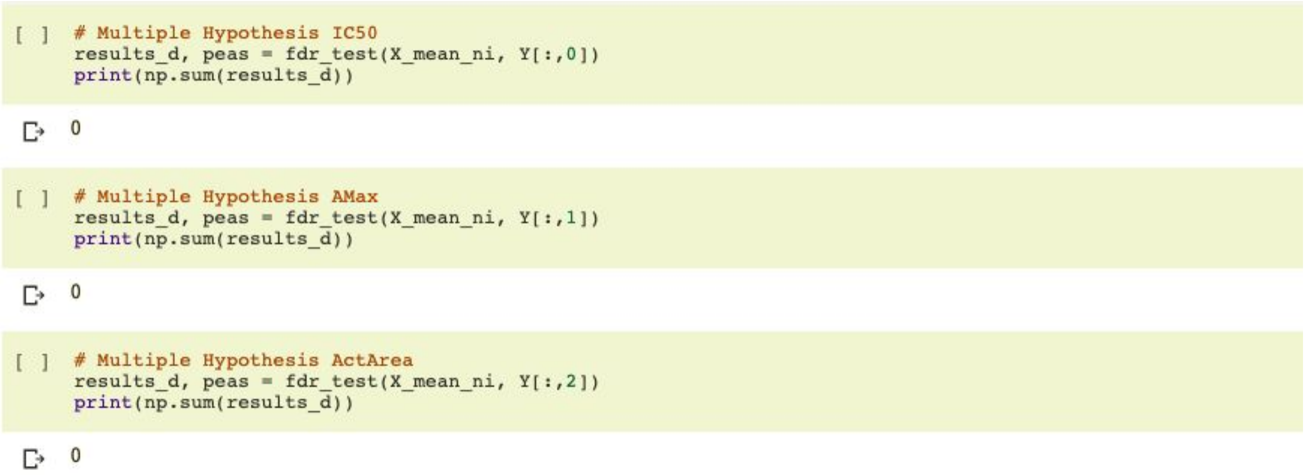


	Lambda	Alpha	MSE Test	MSE Train	R2
118	1.0	1.00	841.694422	792.672591	0.041268
135	3.0	1.00	842.483756	827.473250	-0.000799
117	1.0	0.95	851.585493	800.317666	0.032023
116	1.0	0.90	868.419184	814.514103	0.014848
100	0.3	0.95	880.875913	761.221342	0.079336

A_max

Linear regression on individual features showed no significant correlation

- Multiple hypothesis correction to reduce false discovery rate using Benjamini-Hochberg



```
# Multiple Hypothesis IC50
results_d, peas = fdr_test(X_mean_ni, Y[:,0])
print(np.sum(results_d))
```

```
0
```

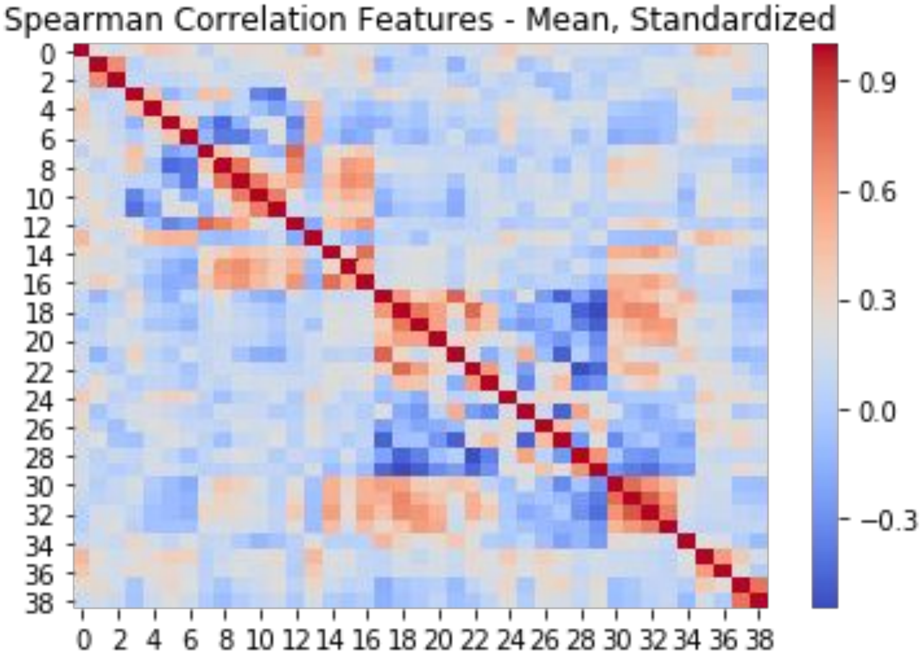
```
# Multiple Hypothesis AMax
results_d, peas = fdr_test(X_mean_ni, Y[:,1])
print(np.sum(results_d))
```

```
0
```

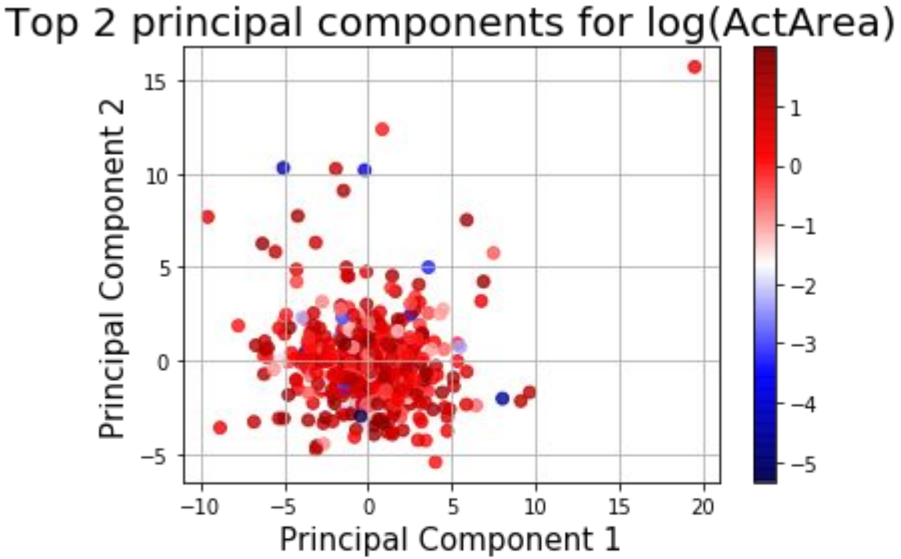
```
# Multiple Hypothesis ActArea
results_d, peas = fdr_test(X_mean_ni, Y[:,2])
print(np.sum(results_d))
```

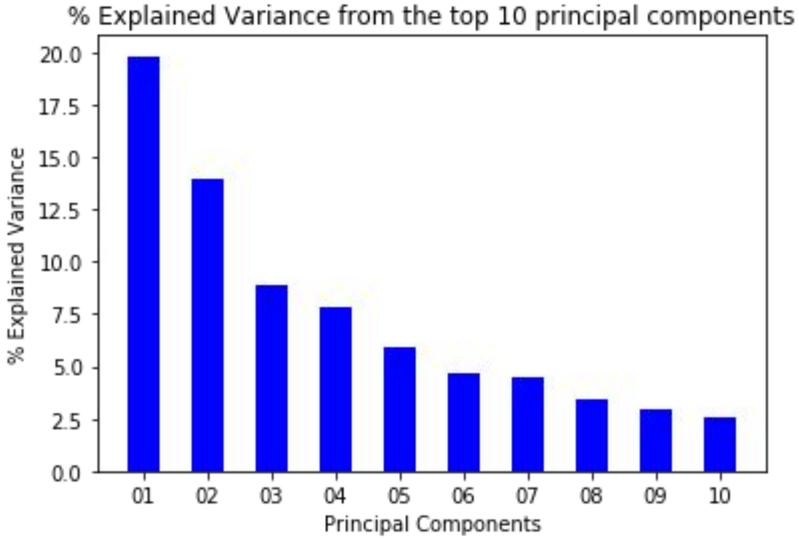
```
0
```

Heatmap shows little correlation between 39 features



PCA shows no distinction between features and data





Absolute Pearson correlation between PCs and features was < 0.1

Reasoning our results

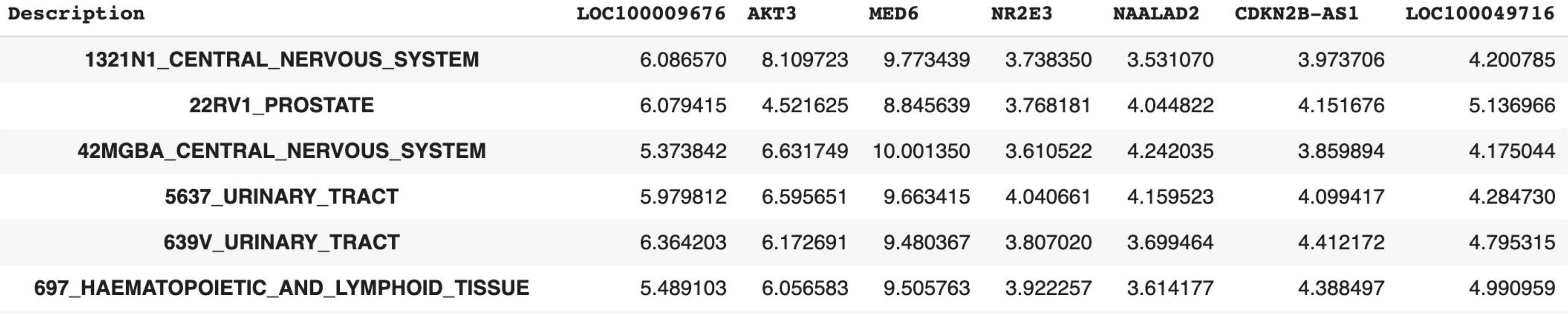
1. May still have correlation
    a. We did not have enough features (39) for the patterns/relations to be captured
    b. Since this is mass spectrometry downstream data, there could have been too many confounding factors that muffled out patterns for a signalling pathway.

2. There could be no correlation.

Gene Expression Analysis

Resolve Missing Values

- Missing values only present in EC50, so we decided not to use it as a label. (134 missing values)
- No missing values in gene expression data



Description	LOC100009676	AKT3	MED6	NR2E3	NAALAD2	CDKN2B-AS1	LOC100049716
1321N1_CENTRAL_NERVOUS_SYSTEM	6.086570	8.109723	9.773439	3.738350	3.531070	3.973706	4.200785
22RV1_PROSTATE	6.079415	4.521625	8.845639	3.768181	4.044822	4.151676	5.136966
42MGBA_CENTRAL_NERVOUS_SYSTEM	5.373842	6.631749	10.001350	3.610522	4.242035	3.859894	4.175044
5637_URINARY_TRACT	5.979812	6.595651	9.663415	4.040661	4.159523	4.099417	4.284730
639V_URINARY_TRACT	6.364203	6.172691	9.480367	3.807020	3.699464	4.412172	4.795315
697_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE	5.489103	6.056583	9.505763	3.922257	3.614177	4.388497	4.990959

•••

Method: Elastic Net Regression

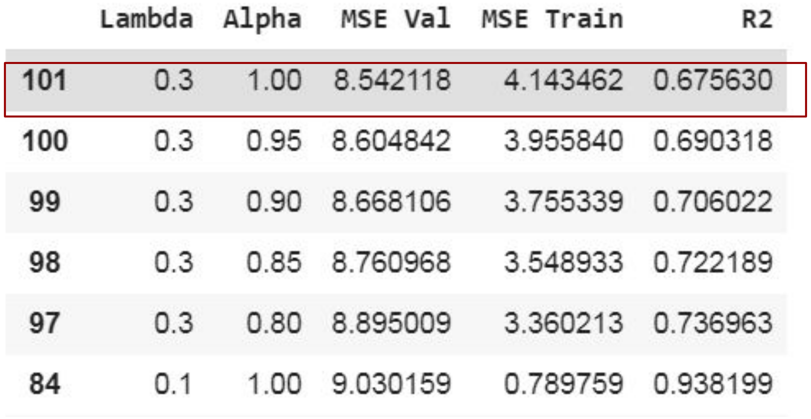
- Linear regression with L1 and L2 regularization
- Supposed to be better at dealing with situations with correlations between parameters
- Loss function:

$$L_{enet}(\hat{\beta}) = \frac{\sum_{i-1}^{n}(y_i - x_i'\hat{\beta})^2}{2n} + \lambda(\frac{1-\alpha}{2}\sum_{j-1}^{m}\hat{\beta}_j^2 + \alpha\sum_{j-1}^{m}|\hat{\beta}_j|),$$

- **α** = mixing parameter (lower → less L1)
- **λ** = regularization parameter (0 = no regularization)

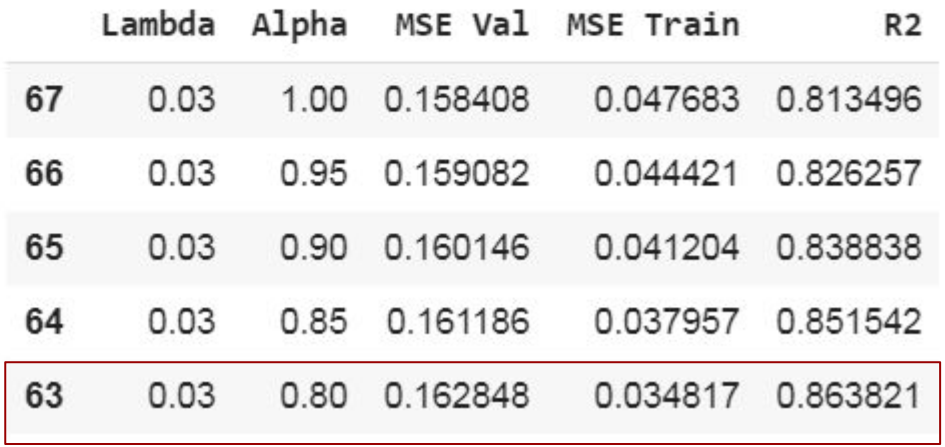
Results

From hyperparameter tuning



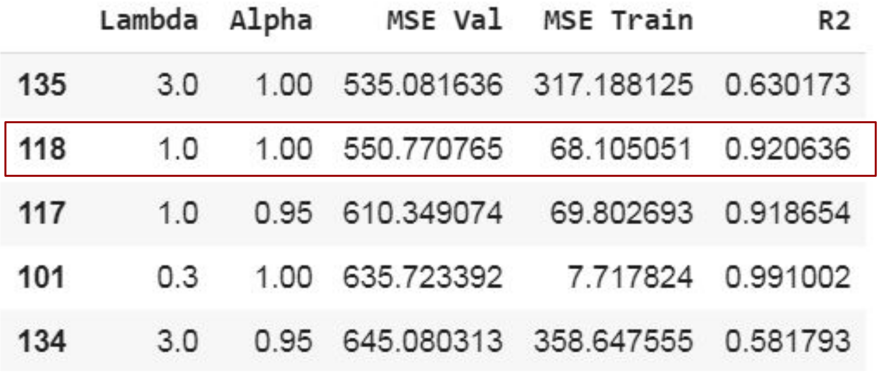
	Lambda	Alpha	MSE Val	MSE Train	R2
101	0.3	1.00	8.542118	4.143462	0.675630
100	0.3	0.95	8.604842	3.955840	0.690318
99	0.3	0.90	8.668106	3.755339	0.706022
98	0.3	0.85	8.760968	3.548933	0.722189
97	0.3	0.80	8.895009	3.360213	0.736963
84	0.1	1.00	9.030159	0.789759	0.938199

IC 50



	Lambda	Alpha	MSE Val	MSE Train	R2
67	0.03	1.00	0.158408	0.047683	0.813496
66	0.03	0.95	0.159082	0.044421	0.826257
65	0.03	0.90	0.160146	0.041204	0.838838
64	0.03	0.85	0.161186	0.037957	0.851542
63	0.03	0.80	0.162848	0.034817	0.863821

Activity Area

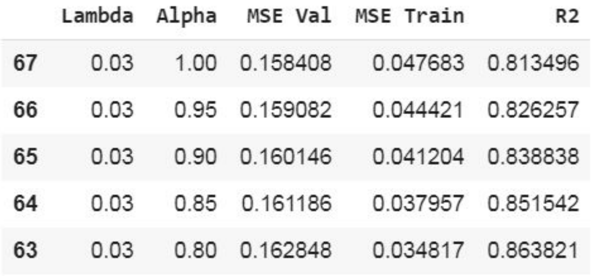


	Lambda	Alpha	MSE Val	MSE Train	R2
135	3.0	1.00	535.081636	317.188125	0.630173
118	1.0	1.00	550.770765	68.105051	0.920636
117	1.0	0.95	610.349074	69.802693	0.918654
101	0.3	1.00	635.723392	7.717824	0.991002
134	3.0	0.95	645.080313	358.647555	0.581793

A_max

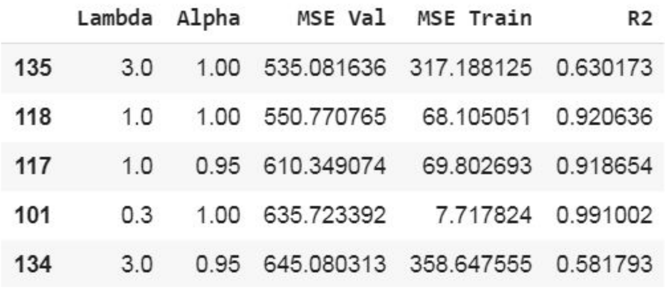
Analysis of hyperparameter tuning

- Paper
  - Did not state their hyperparameter
  - Stated they chose ones with smallest MSE



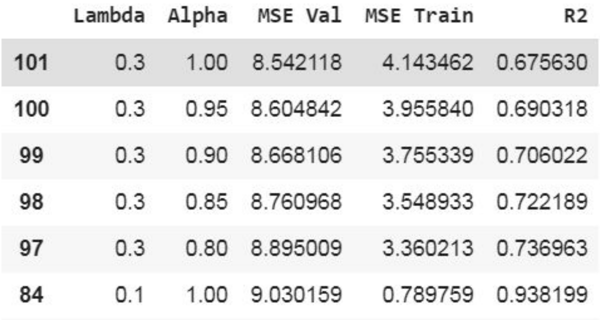
	Lambda	Alpha	MSE Val	MSE Train	R2
67	0.03	1.00	0.158408	0.047683	0.813496
66	0.03	0.95	0.159082	0.044421	0.826257
65	0.03	0.90	0.160146	0.041204	0.838838
64	0.03	0.85	0.161186	0.037957	0.851542
63	0.03	0.80	0.162848	0.034817	0.863821

Activity Area



	Lambda	Alpha	MSE Val	MSE Train	R2
135	3.0	1.00	535.081636	317.188125	0.630173
118	1.0	1.00	550.770765	68.105051	0.920636
117	1.0	0.95	610.349074	69.802693	0.918654
101	0.3	1.00	635.723392	7.717824	0.991002
134	3.0	0.95	645.080313	358.647555	0.581793

A_max



	Lambda	Alpha	MSE Val	MSE Train	R2
101	0.3	1.00	8.542118	4.143462	0.675630
100	0.3	0.95	8.604842	3.955840	0.690318
99	0.3	0.90	8.668106	3.755339	0.706022
98	0.3	0.85	8.760968	3.548933	0.722189
97	0.3	0.80	8.895009	3.360213	0.736963
84	0.1	1.00	9.030159	0.789759	0.938199

IC 50

Bootstrapping for finding significant features

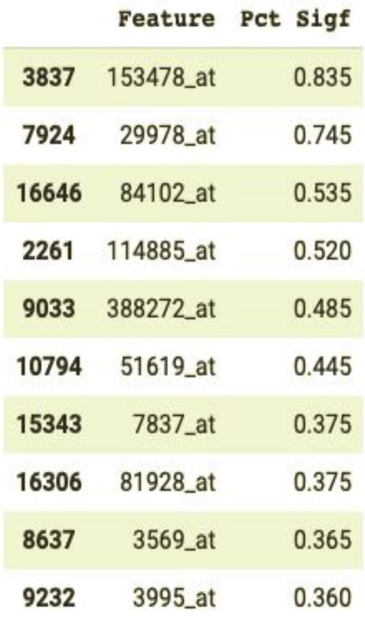
- Generated 200 resampled datasets by sampling train data with replacement
- Each consists of 250 samples, about (1-1/e) or 63% of the training data size
- Solve elastic net for each bootstrap dataset
  - Used the optimal α and λ from the hyperparameter training
- Generate a matrix of regression coefficients β
  - Each row, *k*, represents the solution for one bootstrap dataset
  - Each column, *j*, is the weight on that feature
- Calculate percentage of bootstrap datasets inferred as significant for each feature:

$$r_j = \sum_{k=1}^{200} \left(1_{\setminus 0}\left(\beta_{j,k}^{BS}\right)\right) / 200$$

$1_{\setminus 0}$ is the indicator function

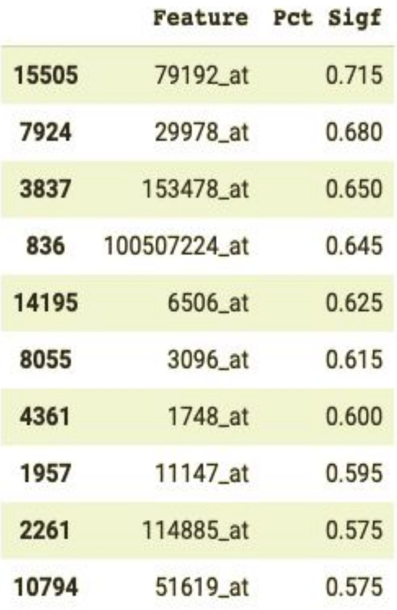
$$1_{\setminus 0}(x) = \begin{cases} 0 \text{ if x=0,} \\ 1 \text{ otherwise.} \end{cases}$$

Top 10 genes that showed up frequently during the bootstrap



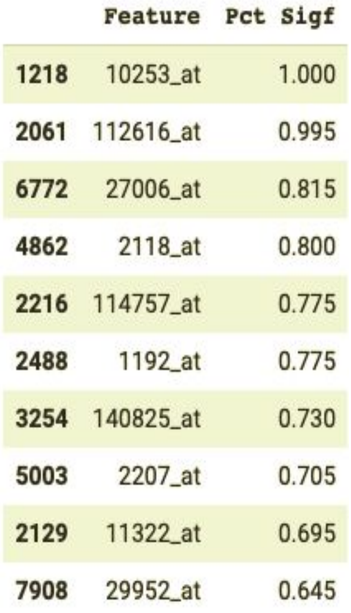
	Feature	Pct Sigf
3837	153478_at	0.835
7924	29978_at	0.745
16646	84102_at	0.535
2261	114885_at	0.520
9033	388272_at	0.485
10794	51619_at	0.445
15343	7837_at	0.375
16306	81928_at	0.375
8637	3569_at	0.365
9232	3995_at	0.360

IC50



	Feature	Pct Sigf
15505	79192_at	0.715
7924	29978_at	0.680
3837	153478_at	0.650
836	100507224_at	0.645
14195	6506_at	0.625
8055	3096_at	0.615
4361	1748_at	0.600
1957	11147_at	0.595
2261	114885_at	0.575
10794	51619_at	0.575

A_{max}



	Feature	Pct Sigf
1218	10253_at	1.000
2061	112616_at	0.995
6772	27006_at	0.815
4862	2118_at	0.800
2216	114757_at	0.775
2488	1192_at	0.775
3254	140825_at	0.730
5003	2207_at	0.705
2129	11322_at	0.695
7908	29952_at	0.645

Act Area

Top 10 genes that showed up frequently during the bootstrap for Activity Area

Gene ID	Gene Name	Function	Frequency	Ave Weight
10253_at	SPRY2	Regulator of MAPK output	1	0.655
112616_at	CMTM7	Tumor suppressor	0.995	0.728
27006_at	FGF22	Mitogenic and cell survival activities	0.815	0.464
2118_at	ETV4	RET Signaling	0.8	0.723
114757_at	CYGB	Protective function during oxidative stress	0.775	0.697
1192_at	CLIC1	Chloride intracellular channel	0.775	0.445
140825_at	NEURL2	Regulation of myofibril organization	0.73	0.482
2207_at	FCER1G	IgE receptor involved in allergic reactions	0.705	0.649
11322_at	TMC6	High rate of progression to squamous cell carcinoma	0.695	0.492

Weights have flipped signs when compared to the paper’s top features

		Our model		Barretina et al.	
Gene Name	Function	Frequency	Ave Weight	Frequency	Ave Weight
SPRY2	Regulator of MAPK output	1	0.655	0.980	-0.328
NEURL2	Regulation of myofibril organization	0.73	0.482	0.845	-0.120
CMTM7	Tumor suppressor	0.995	0.728	N/A	-0.014
FCER1G	IgE receptor involved in allergic reactions	0.705	0.649	N/A	-0.008

Differences between our model and the paper

- Log-transformed labels
- SNP and CNV data not included in our model

Results from bootstrapping:

- Train MSE: 0.039 – Test MSE: 1.63

Conclusions

- No correlation was found between global chromatin profiling data has with MEK inhibitor sensitivity
- Gene expression data can be used for prediction of MEK inhibitor's activity area
- Our top bootstrapped feature is the paper's top bootstrapped feature for PD-0325901, but flipped in sign

Thank you! – Questions?

References

Gillet, J. P., Varma, S., & Gottesman, M. M. (2013). The clinical relevance of cancer cell lines. Journal of the National Cancer Institute, 105(7), 452-458.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., ... \& Reddy, A. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature, 483(7391), 603.

Litichevskiy, L., Peckner, R., Abelin, J. G., Asiedu, J. K., Creech, A. L., Davis, J. F., ... \& Jaffe, J. D. (2018). A library of phosphoproteomic and chromatin signatures for characterizing cellular responses to drug perturbations. Cell systems, 6(4), 424-443.

Jaffe, J. D., Wang, Y., Chan, H. M., Zhang, J., Huether, R., Kryukov, G. V., ... \& Stegmeier, F. (2013). Global chromatin profiling reveals NSD2 mutations in pediatric acute lymphoblastic leukemia. Nature genetics, 45(11), 1386–1391. doi:10.1038/ng.2777

Rahman, R., \& Pal, R. (2016, February). Analyzing drug sensitivity prediction based on dose response curve characteristics. In 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI) (pp. 140-143). IEEE.

Data retrieved from: https://portals.broadinstitute.org/ccle

https://www.diagenode.com/en/categories/histone-antibodies

https://www.researchgate.net/figure/Schematic-representation-of-histone-modifications-The-methylation-sites-are-represented_fig1_283086163