Predicting Cytokine Responses From Image Embeddings

Tempest Plott 2023-09-18

Problem Statement

The Question:

The two questions addressed by the analysis presented in this report are -

1) Can the production of any of these 51 cytokines be predicted from image embeddings alone?

and

2) Which feature selection method is the most successful for these predictions?

For this analysis, I define predictable cytokines as those whose production can be predicted with at least 0.65 precision and 0.65 recall under at least one feature selection strategy. (In this analysis, chance would result in scores of 0.5.) The ability to predict numerous cytokines will also be considered in the successfulness of a feature selection strategy. Finally, faster feature selection strategies will be preferred because this method will be scaled up and productionized if prediction is possible.

There is no expectation that any of these cytokines will be predictable with the given data, so this analysis is an initial proof-of-concept and model improvement should be performed in follow-up work.

These questions require more context themselves, so please read the following background information section if you are curious about it.

Background Information:

Cytokines are chemical messages sent between cells to coordinate activities. Many cytokines have been well-characterized, such that production of certain cytokines is known to indicate specific cellular activities. For example, scientists can know which immune cells have become activated after a stimulus by measuring specific cytokine levels. To collect and analyze cytokines, liquid is carefully removed from the top of incubating cells, frozen, and generally shipped to a third party. This is often a risky, expensive, and time-consuming process.

High-content imaging and machine-learning based analysis is quickly becoming the standard method of screening for new drugs. With HCI and ML, scientists can rapidly screen hundreds of thousands of novel drugs and compare their similarity to controls, hunting for drugs which match a desired effect or reverse an unwanted effect. However, with similarity scoring alone, scientists cannot understand the reasons for the similarity to a control. Two images might appear very similar for completely different biological reasons. So, to move drug candidates forward, additional cytokine measurement data is often needed.

If it can be shown that image embeddings can predict cytokine production, then the relationship between images and known cytokine context can be put together so more cures can be found more easily and more quickly. That is the goal of this proof-of-concept analysis.

Data Collection

384 wells each containing 18,000 human white blood cells were plated.

51cytokines were quantified for each of these wells. Each cytokine measure was expressed as log10-fold change relative to the mean of negative control wells and median-shifted to center at 0. This procedure is important for consistent communication between scientists and was therefore performed before handing off the data for further analysis.

Each well was imaged with Concanavalin-A, a fluorescent dye which stains cell membranes. These microscopic images were analyzed with a CNN to generate 1230 black box features for each cell in the image. This analysis does not cover the process of generating those embeddings, which had already been chosen and performed by the client. Briefly, the embeddings were obtained with an open source CNN called EfficientNetV2XLImageNet21 for every single-cell crop, z-score normalized per plate and donor, and uploaded in a parquet file for this work. This embedding process creates 1230 features for each of five fluorescent channels for every cell imaged in order to capture the rich information of the images. That is why the feature space must be reduced to scale up the analysis in the future. Because Concanavalin-A shows the overall shape, size, and to some degree texture of the cell, it is a good generalized dye to start this analysis with.

Cleaning the Data

The embedding features were then aggregated by well (via the mean), since cytokine measurements are necessarily aggregated by the well and cannot be collected for each individual cell.

Each of 51 cytokine distributions was transformed into two bins with KBinsDiscretizer and labeled with 1 or 0 to indicate each well as either being a producer (1) or a non-producer (0) for each cytokine. Cytokines that had the same level of production in every single well were removed from the experiment. If a cytokine shows the same exact number for the entire experiment, that realistically means the measurement was not successful, likely due to the true values being outside of the range of the assay. (eg, that number is simply the floor or ceiling of the assay.) 40 cytokines survived this process. See Figures 1 and 2 below.

Figure 1: Histogram of experiment-wide normalized MMP-9 values.

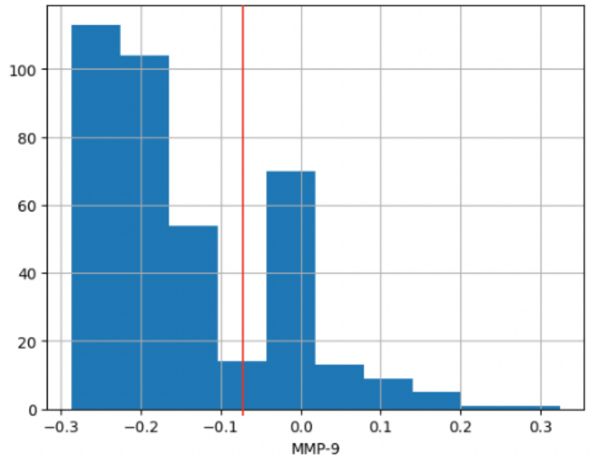


Figure 1: Example of KBinsDiscretizer splitting cytokine measurements into two bins. The Y axis is the number of wells and the X axis is the normalized cytokine value.

Figure 2: Histogram of experiment-wide normalized TGF-beta 1 values.

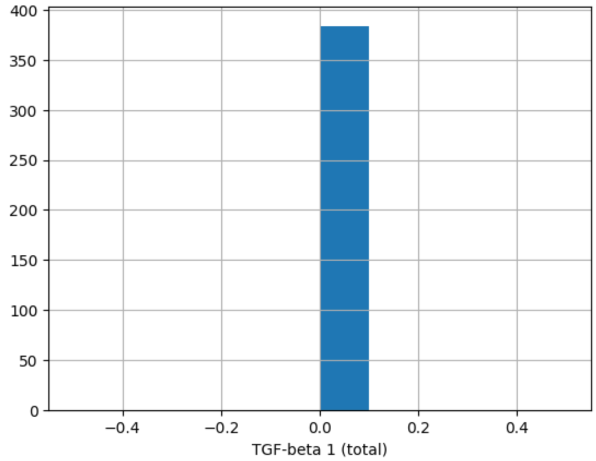


Figure 2: Example of a rejected cytokine. The Y axis is the number of wells and the X axis is the normalized cytokine value.

Exploratory Data Analysis

The distributions of binarized cytokine production were observed. It was noted that there are many more non-producing wells than producing wells for essentially all cytokines, so a downsampling strategy was employed per-cytokine to remove non-producing wells at random to match the number of producing wells. This perfectly balanced the classes for the later modeling steps. Cytokines which had fewer than 10 wells after this downsampling process were dropped from the experiment. This left 26 cytokines to analyze.

Figure 3: Histogram of experiment-wide binarized cytokine labels.

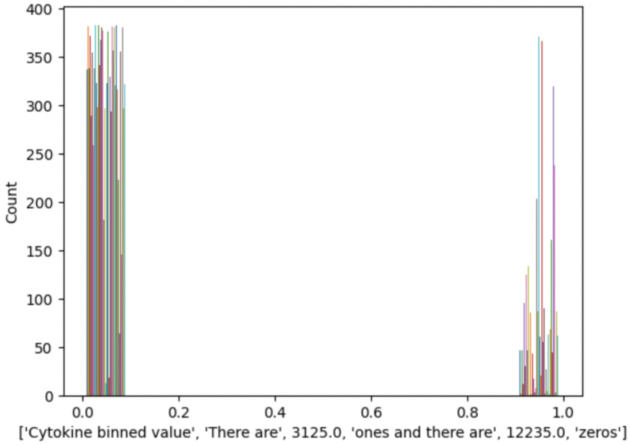


Figure 3: There was about a 4:1 ratio of wells labeled 0 to wells labeled 1 before downsampling 0s for each cytokine at random.

Data Wrangling

A dataframe with all 1230 features and balanced classes was thus constructed for each cytokine. These dataframes were stored in a list so that the list could be looped through in feature selection and modeling. In other words, each cytokine has its own model. It is important to note that some of the cytokines are independent of each other, while others are naturally produced in concert with each other. It was thus important to separate each cytokine into its own model to see if the final XGBoost classifier (and feature selection methods that use classifiers) can truly learn to predict just that cytokine rather than “cheating” by using the signal from another cytokine as an input. Thus, it was also important to downsample and pre-process each cytokine separately. Proper data types were also enforced at this stage.

Feature Selection

Seven feature space reduction techniques were compared. These are:

1. Variance thresholding (Discard features with low variance.)
2. Univariate Feature Selection (f_regression: Use the F value to keep all features statistically correlated with the labels.)
3. L1 feature elimination (LASSO regularization: Keep all features strongly positively correlated with the labels.)
4. Tree-based feature selection (Create an initial decision tree and remove non-useful features as defined by Gini purity.)

5. Sequential feature selection (Create a logistic regression model for all features individually, choose the best single feature, and repeat this process slowly adding one feature at a time until a desired number of features is reached.)
6. Recursive feature elimination (Starting with all features, repeatedly construct an SVM model, determine feature importance, and remove the least important feature until a desired number of features is achieved.)
7. PCA (This method is more like feature engineering. It transforms the feature space into a desired number of dimensions on axes that best explain the variance of the data.)

For those methods which require the user to specify a number of features, 10 features were always chosen. Interestingly, no particular blackbox feature or group of features was repeatedly selected. In other words, the different feature selection methods all focused on different features as key and there was no superstar feature.

To compare these methods fairly, each downsized feature space was used in the same XGBoost model architecture. Specifically, model = XGBClassifier(n_estimators=100, max_depth=10, learning_rate=1, objective='binary:logistic'.) The train/test split was always 50/50, giving the comparison model a generous chance to learn much from the feature space.

For those methods which naturally generate a different number of features for every dataset they are used on, analysis was performed to see whether the classifier model would score better simply based on the number of features. There was no trend showing that the number of features alone improved the model. Please see an example in Figure 4.

Figure 4: A larger number of features does not correlate with better precision or recall for Univariate Feature Selection.

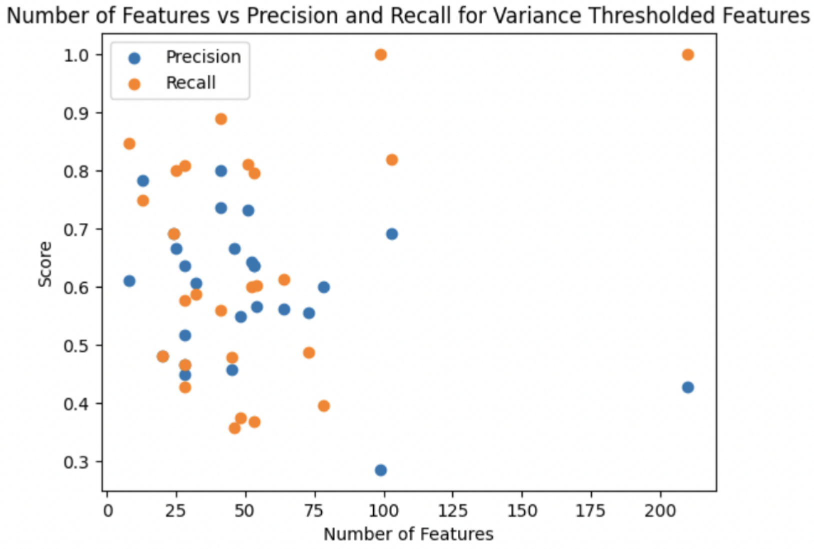


Figure 4: Example of a feature selection method which generates a variable number of features. More features does not necessarily lead to higher precision or recall. Here, each cytokine has a different number of features.

Similar analysis was performed to see if cytokines with more wells that survived the downsampling process generally performed better. There was no trend observed that the number of wells alone improved the model under any of the feature selection strategies. Please see Figure 5 for an example.

Figure 5: A larger test set does not correlate with better precision or recall for Recursive Feature Elimination.

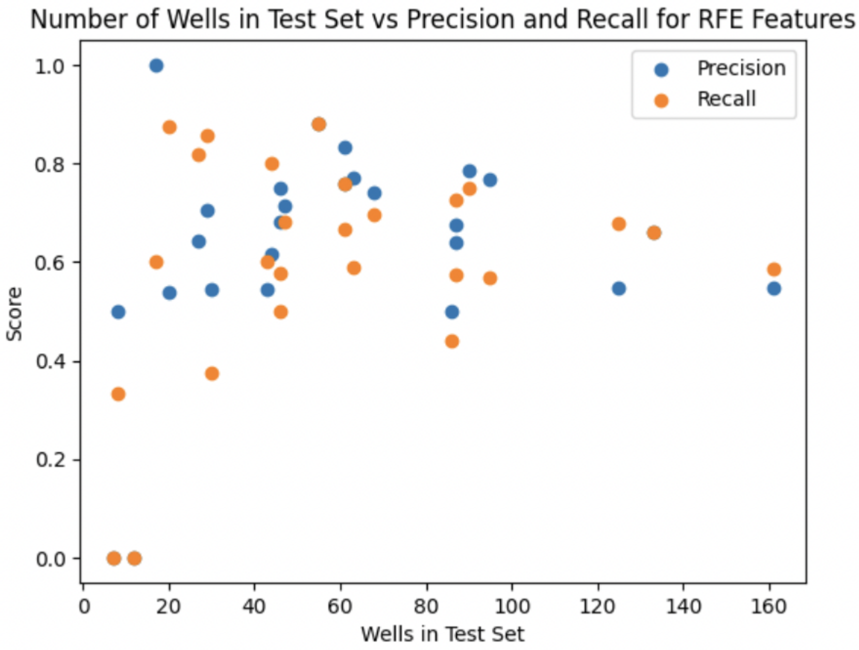


Figure 5: Here, each cytokine has a different number of wells in its test set. Simply having a larger test set does not correlate with improved performance.

Results

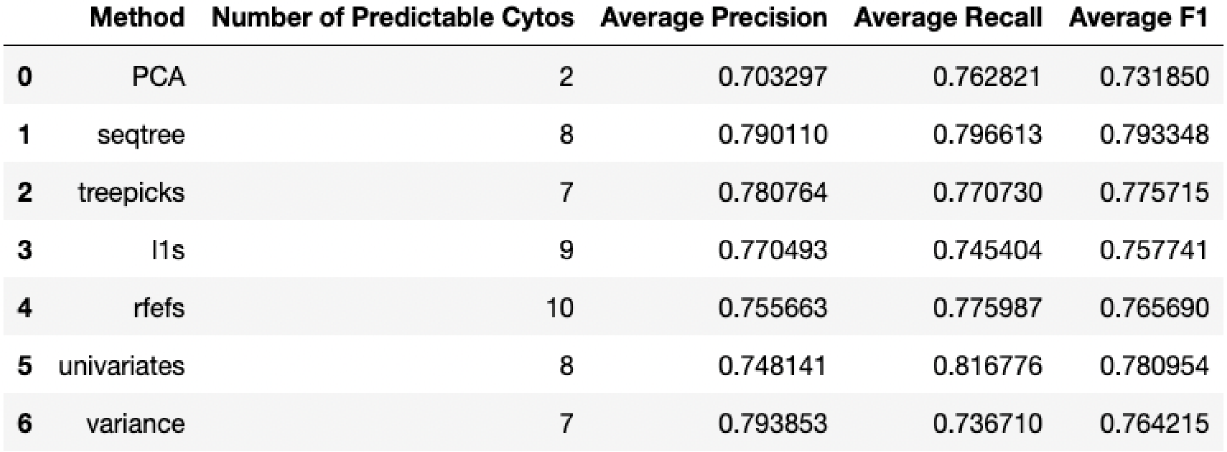
The answer to the question “can any cytokine responses be predicted from ConA embeddings?” is a resounding “yes!” Even with a yet-unoptimized model and only one fluorescent channel, several cytokines had precision and recall scores better than 0.65 for numerous features selection strategies. Some cytokines appear to be more easily predictable than others.

Figure 6: Cytokines and the number of feature selection methods (out of seven) which they were successfully predicted with.

	G-CSF	IL-1 beta	IL-16	CCL20	GM-CSF	CCL1	CD14	CCL19	CXCL10	IL-1 alpha	Activin A	IL-6 R alpha	CCL2	IL-12 p40	CRP	MMP-9	IL-8	CXCL5
count	6	6	5	5	4	4	3	3	3	2	2	2	1	1	1	1	1	1

As promised, each feature set was then compared based on its overall performance as defined by precision, recall, and the number of predictable cytokines generated.

Figure 7: Results of Feature Selector Comparison, by Feature Selector Nickname

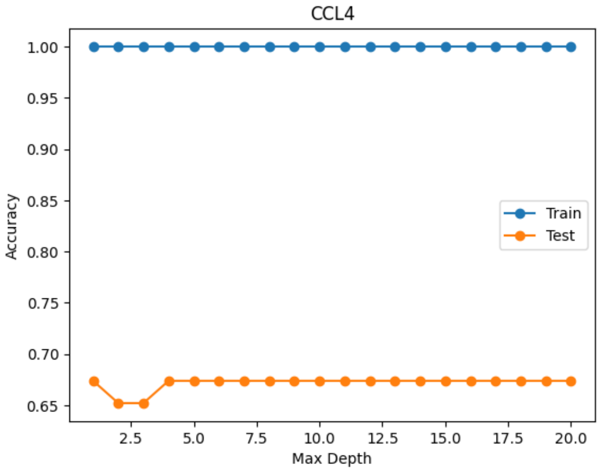


	Method	Number of Predictable Cytos	Average Precision	Average Recall	Average F1
0	PCA	2	0.703297	0.762821	0.731850
1	seqtree	8	0.790110	0.796613	0.793348
2	treepicks	7	0.780764	0.770730	0.775715
3	l1s	9	0.770493	0.745404	0.757741
4	rfefs	10	0.755663	0.775987	0.765690
5	univariates	8	0.748141	0.816776	0.780954
6	variance	7	0.793853	0.736710	0.764215

It is a close race between these feature selection methods. Forward sequential feature selection technically had the highest scores as shown here, but not the most predicted cytokines, and is stymied by the fact that even one channel needs massive computing power to go through that process. On my local machine, the method took four hours to run for one channel on one plate of data. Therefore, recursive feature elimination is the best method in my opinion. It runs relatively cheaply (about three minutes on my machine,) so it can easily be scaled up to five channels. I would recommend taking the top ten features for each channel and combining them for the final model.

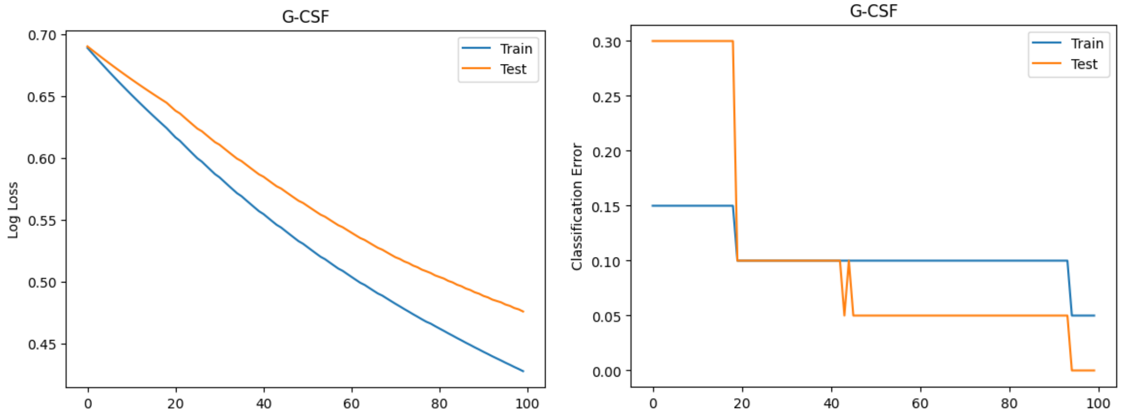
The model was thus further evaluated with the feature set generated by Recursive Feature Elimination. A grid search was performed for the max depth of the XGBoost model, and there was a general trend across cytokines that a depth of more than 3 was not necessary and often harmful. Please see the example in Figure 8 below.

Figure 8: Example of Effect of Max Tree Depth on Performance of Predicting a Cytokine



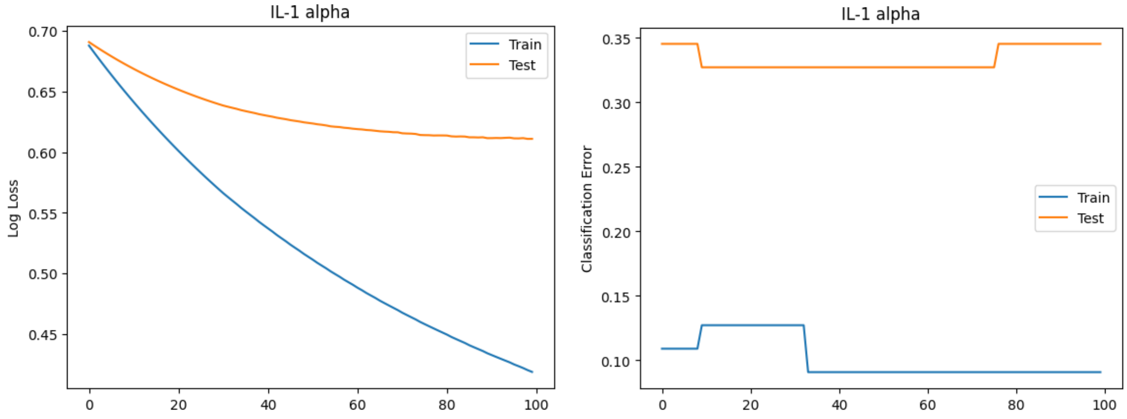
Next, a grid search of early stopping points was used to check for overfitting. In this analysis, the logloss and Classification Error are plotted against the number of epochs used for the train and test sets. If the test set error does not decrease like the training set, that is indicative of overfitting to the training set. The performance of G-CSF, one of the most easily predicted cytokines, is shown in Figure 9.

Figure 9: The G-CSF model is not overfit.



However, this good news is contrasted by the performance of IL-1 alpha which was also predictable with the Recursive Feature Elimination feature set. The test set performance does not improve alongside the training performance as the number of epochs increases. There was no general stopping point which would be beneficial to all cytokines. Thus, each model needs to be optimized independently.

Figure 10: The IL-1 alpha model is overfit.



Future Research

Clearly, then, a point of further research would be full parameter sweeps to improve the ultimate classification model for each cytokine. This should be performed with all five channels.

Feature importance analysis should be done to see which channel (ie, which organelle of the cell) is most informative for predicting any particular cytokine response. That way, the laboratory will know which channels are important depending on the particular cytokine of interest.

Also, if the fine differences in performance of each feature selection method are important, this analysis should be run many times and the scores collected to show which feature selection method truly has the best score and whether they are even significantly different from each other in that respect.

Another follow up question for the scientists is to comment on the identity of the cytokines which are predictable here. Does it fit their domain knowledge of which cytokines would likely be predictable from ConA alone?

This work can be used as 1) a proof of concept which validates the ability of the current embeddings to capture real biological signal, 2) an initial model which can be used to predict some cytokines from PBMC ConA data, and 3) a valuable note that the embeddings are rich and many feature selection methods of them will perform well, so the most convenient ones for the company can likely be chosen in other analysis.

Thanks

Gratitude to Noor Hussain and Ahmed Hosny for their guidance in performing this analysis.

Three outside resources were used in this work.

This page by sklearn lists the feature selection strategies used in this notebook, with the addition of PCA.

Code from this post was used/adapted to create more visually appealing confusion matrices.

Code from this post was used/adapted to show learning curves and look for evidence of overfitting.