CTD2BeatAML DREAM Challenge: A Multi-task Model for Predicting Drug Responses

Ruibo Zhang, Daniel Nolte, Saugato Rahman Dhruba

\*Affiliations\*

Abstract

Our model utilizes multi-task bagged linear regression with post-prediction covariance alignment, to predict the AUC values of individuals inhibitor responses.

Introduction

In sub-challenge 1, the AUC values of tissue specimens under 122 inhibitors are given for training and are the target to be predicted. In Figure 1, the AUC values are plotted vs. their ranking within each of the selected inhibitors, after the mean AUC values are subtracted from the initial AUC values. From the figure, we did not observe a linearity in the distribution of the AUC values: the AUC values have significant non-linear tails, meaning the highly responsive and highly resistive specimens exhibit extreme AUC values. This inspired us to use the normalized ranking of the AUCs, instead of the real AUC values, to better utilize a linear model.

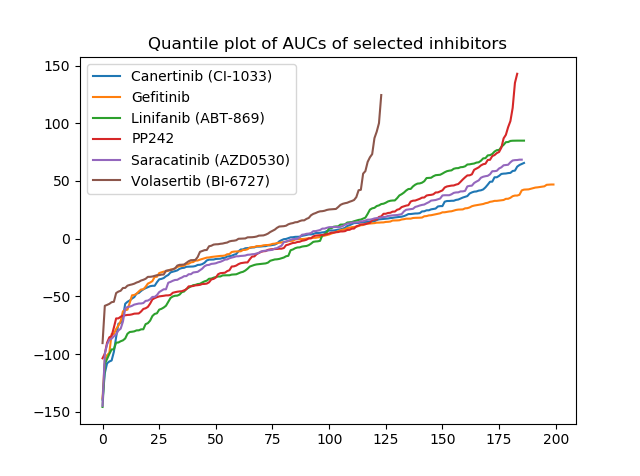


Figure 1, distribution of the AUC values of selected inhibitors shown in order.

Cancer drug design usually follows the logic of pathway control and relationships between chemical groups and biochemical reactions. Therefore, similar inhibitors are supposed to have similar responses with similar cell specimens. This can be represented by the correlation / covariance structures between the AUCs of different inhibitors. The correlations between the AUC values of the 122 inhibitors are shown in Figure 2. Inhibitors are first clustered according to the correlation distances and shown in order of the clustering. As shown in Figure 2, the correlations between inhibitors show a clear pattern, where clustered inhibitors have correlation coefficients as high as 0.7 to 0.8, while uncorrelated inhibitors have coefficients as low as -0.3 to -0.4.

It is reasonable to assume that the predicted AUC values of all drugs should preserve the covariance structure similar to that of the training data. To pass the original covariance structure to the predictions, we adopted a covariance alignment method, CORAL, which is well accepted in the cross-domain learning[[1]](#endnote-1). After obtaining the predictions of all inhibitors, we divide the predictions by the square root of its covariance matrix to remove its own covariance structure, which is called “whitening”. We then multiply the whitened prediction with the square root of the covariance of the training data, which is called “re-coloring”. From another aspect, we are meanwhile minimizing the higher order prediction residuals (covariance, in order of 4), besides the MSE in linear regression, which is in the order of 2.

To utilize the multi-task intrinsic qualities of our task, bagged linear regression is used as the base model, which is compatible with multi-task regression. Bagging compensates the instability of simple linear regression and provides improved prediction accuracy.

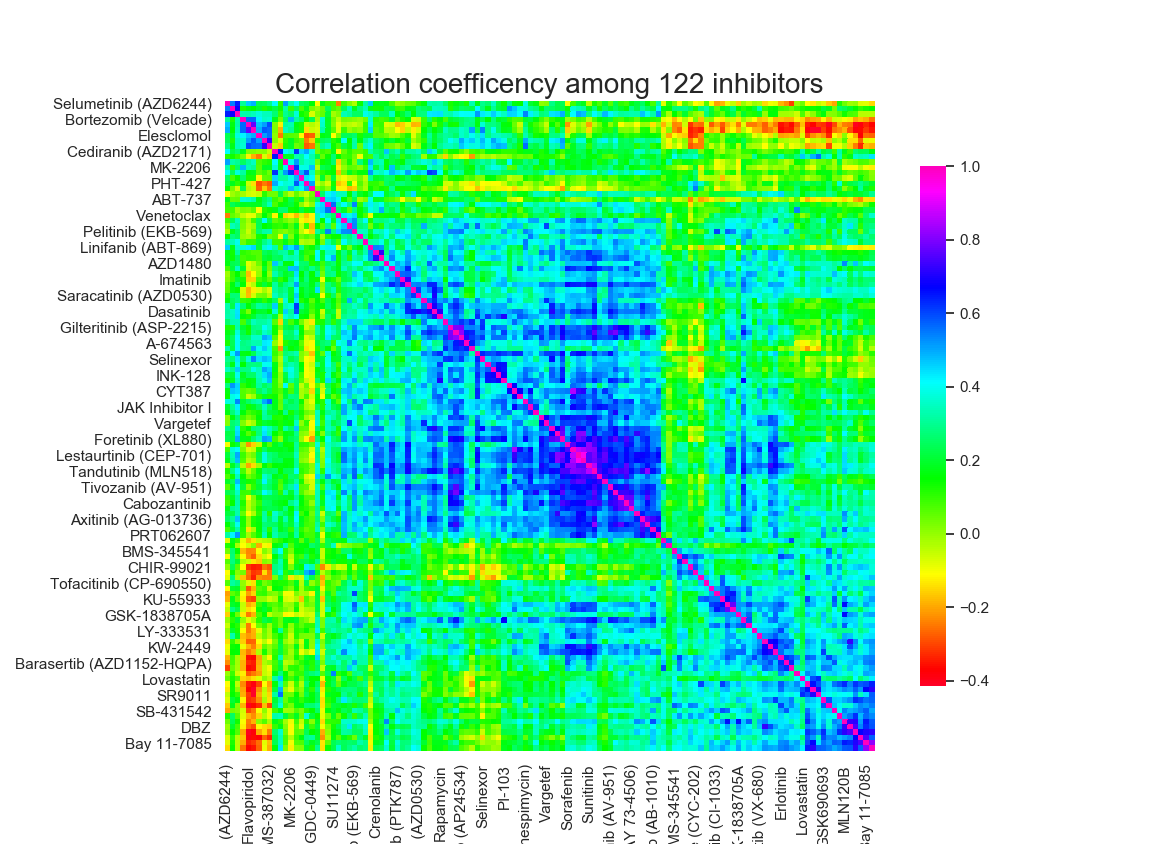


Figure 2. Correlation ecoefficiency among 122 inhibitors.

Methods

Data preprocessing

The training set and leaderboard set are both preprocessed in the way described below. In the SC1, RNA sequences, DNA sequences, and clinical data are provided as input for prediction. In our model, only RNA expressions are used. First, the RNA expressions are normalized by dividing the 2-norm of each specimen, followed by the feature selection process. Then RNA expressions are standardized so that the mean of each gene is 0 and std is 1 before modeling. AUCs are given as target values. Not all inhibitors are tested on every specimen, in our case, specimens with more than 30 NaNs are dropped out, and the remaining missing values are imputed using KNN imputation, with k = 5. Instead of using the real values directly, we used the normalized AUC rankings of the specimens as the target. The mean of the normalized AUC ranking of each inhibitor is subtracted from the target values before model fitting.

Feature selection

In the given dataset, 63677 genes are supplied. First, the RNA dataset is filtered to keep the 30000 genes with largest variances. Then a ridge regression model is fitted, and the features with the highest weights are kept. In our case, ~ 19 – 20k features are kept and put into the bagged linear regression model for training.

Regression model

A bagged linear regression estimator is used as the prediction model. To avoid overfitting, only a portion of features are kept for training each single linear regressor. Less than 100 features are selected to put into each of the 5000 linear regressor’s each time.

Post processing

To reduce the covariance discrepancy between training AUCs and predictions, we used the method of covariance alignment (CORAL) described by Sun et.al[[2]](#endnote-2). For efficiency, stability, and to avoid failures in taking the inverse of the covariance matrix, we simply add a regularization term \*lambda I\* to the covariance matrices. First predictions are divided by the square root of its covariance and multiplied by the square of training data covariance. This can be described as:

where is the regularized prediction covariance, tr is the regularized training AUC covariance, and is the pristine prediction.

In our project, =1 was used.

Discussion

As shown in Table 1: 9 models are evaluated by 5-fold cross-validation using both training and leaderboard data. First, Ridge regression and bagged linear regression on actual AUC values are shown as the baseline. AUC rankings are used instead of AUC values in models #3 and #4. Further, post-prediction alignment is shown as model #5 - 8. As a comparison, Bagged LR with Ridge feature selection but not post-processing is given as #9.

Table 1. Summary of model performances.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Training values | Features kept in 1st feature selection | the 2nd feature selection | Post processing | Regression model | Model parameter | Spearman correlation | Pearson correlation |
| #1 | values | 30k | - | - | Ridge | alpha=1 | 0.322344 | 0.318868 |
| #2 | values | 30k | - | - | Bag LR | Max feature portion=0.005 | 0.32371 | 0.312633 |
| #3 | ranks | 30k | - | - | Ridge | alpha=1 | 0.327609 | 0.314593 |
| #4 | ranks | 30k | - | - | Bag LR | Max feature portion =0.005 | 0.328202 | 0.314058 |
| #5 | ranks | 30k | - | CORAL | Ridge | alpha=1 | 0.336019 | 0.323188 |
| #6 | ranks | 30k | Ridge | CORAL | Ridge | alpha=1 | 0.330447 | 0.320249 |
| #7 | ranks | 30k | - | CORAL | Bag LR | Max feature portion =0.005 | 0.333266 | 0.323085 |
| #8 | ranks | 30k | Ridge | CORAL | Bag LR | Max feature portion =0.005 | **0.34171** | **0.326697** |
| #9 | ranks | 30k | Ridge | - | Bag LR | Max feature portion =0.005 | 0.333703 | 0.317601 |

References

1. Sun B., Saenko K. (2016) Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In: Hua G., Jégou H. (eds) Computer Vision – ECCV 2016 Workshops. ECCV 2016. Lecture Notes in Computer Science, vol 9915. Springer, Cham [↑](#endnote-ref-1)
2. Sun, Baochen, Jiashi Feng, and Kate Saenko. "Return of frustratingly easy domain adaptation." *Thirtieth AAAI Conference on Artificial Intelligence*. 2016. [↑](#endnote-ref-2)