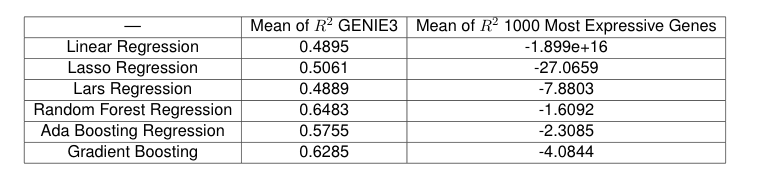
For this milestone 2, we approached the task with the key question of what could be good predictors of the gene expression values of a target gene or a subset of target genes. We structured our approach as such:

* Feature selection and prediction
  + Feature selection methods: GENIE3, most expressed genes, WGCNA, Landmark 1000 genes (L1000)
  + Prediction methods: Linear regression, lasso regression, lars regression, random forest regression, ada boosting regression, gradient boosting
  + Different feature selection methods were matched with different prediction methods to various degrees of completion
* Imputation
  + Methods: naive mean imputation, kNN, variational autoencoder (VAE)

For our regression analysis, we decided to select the features for the gene predictions in two ways: 1) Getting the Coefficient of Variation of all the genes, sort them and choosing the top 1000 genes and use them as our features for all gene predictions. These would be the 1000 most expressive genes by Coefficient of Variation. With these features, we would predict the next 1000 most expressive genes. 2) Using GENIE3, calculate the pairwise correlation of genes. For each gene, it was decided to use as features the ones that have 0.01 or above correlation with it. Due to computational power limitation, only 4000 genes were chosen for this analysis. The results for each one of the models used are shown in the table below:



Even though using the 1000 most expressive genes by Coefficient of Variation allow us to use the same features for all predictions, results were significantly worse than in GENIE3. Nevertheless, a major drawback of the GENIE3 method is that we need the entire database and significant computation power to calculate the pairwise correlation. Given these results, we decided to pursue several more feasible feature selection methods and models that will likely provide better performance

Imputation using kNN and VAE actually achieved good results with R^2 around the range of 0.8-0.9 with 5-50% of the gene expression values missing. Despite the good results, we are hesitant to move forward with imputation methods because they fundamentally impute missing data from the majority data and neglect individual perturbations. This means that it is fundamentally unsuitable for our goal of inferring gene expressions from perturbations. However, the relative success of imputation methods demonstrates that there exist underlying structures in the expression and we have been able to capture them.

One of the main approaches we attempted was a replication of the existing L1000 landmark pipeline study done in human cells, as their methodology and goals closely aligned with our problem statement. Our results and work in this led to a recreation of their processing pipeline for the selection phase, and a number of interesting findings regarding the viability of the approach. While we were successful in recreating their code (from an unusable state that they provide publicly), two primary unique challenges pertaining to our data problem presented themselves:

1. **Lack of perturbations and variability in our dataset.** The existing L1000 selection method relies heavily on clustering with the assumption that variability exists, and uses an aggressive iterative peel-off process that works in a context of being able to throw away entire clusters of genes identified midway, which is not feasible when our results return clusters of the size of thousands.
2. **Computational complexity.** Unlike the paper’s consideration of only 10,000 human genes, our fundamental unit of computation of a 40,000x40,000 genes is 16 times larger, and a standard floating point valued matrix weighs in at **25.6GB** of data, which well exceeds standard laptop and even most cloud computing instance memory. This significantly hampers both computational speed and the types of algorithms that are even available, as working memory sets for mid-algorithm computation can reach unfeasible levels with naive implementations, potentially requiring custom versions or parallelized approaches.

We have the following next steps in mind:

* GENIE3 selection and prediction pipeline: parallelize computation of all 40,000 genes
* L1000 pipeline: waiting for computation to finish and obtain results
* Graph neural networks: embark on the exploration of GNN methods