AC297r Project: Harvard - Inari

Predicting the Effects of Genetic Perturbation in Maize

Victor Avram, Sergio Miguel Moya Jimenez, M. Eagon Meng, Wenhan Zhang

March 30, 2021

Contents

1 INTRODUCTION AND MOTIVATION

As a plant breeding technology company, Inari's work comprises three high-level stages: (1) using computational methods to build genetic knowledge; (2) genetic editing; and (3) delivery of altered genetic information to specific parts of the plant. Our project sits within the first stage of Inari's work and seeks to answer the question "what are the effects on the rest of the maize genes when we perturb a subset of the maize genes?" Essentially, we hope to construct a genetic network that can act as a look-up table to inform Inari of the genetic effects and side-effects of perturbing particular maize genes, as opposed to only observing unexpected effects later in the breeding process.

We found a close counterpart of our problem for human genetics in the Connectivity Map (CMap) project (Subramanian et al, 2017). With the selected landmark 1,000 genes, the CMap project achieved 81% when predicting the rest of the human genome. However, this close reference also illuminates the limitations of our data and significant challenges we face.

Firstly, our dataset is not perturbation-driven such as that in the CMap. By profiling specific genetic perturbations, CMap was able to link variations in genetic profiles directly to specific genes. Secondly, our dataset of 480 tissue samples across 25 individuals is significantly smaller 12,031 genetic profiles CMap analyzed. Lastly, CMap made inference on 11,350 human genes, while we were initially tasked with predicting the rest of the 46,430 maize genes.

Therefore, with these challenges in mind, our EDA has the goals of firstly, reducing the number of relevant genes to be inferred therefore reducing the prediction task, and secondly, observing patterns and linkages in genetic expressions so as to determine whether there is sufficient data for us to create the maize counterpart of CMap.

2 DATA

2.1 Data Description

The data is comprised of gene expression values derived from 26 individual maize plants. Multiple samples were taken from 10 different tissues from these 26 individuals and subsequently sequenced in order to obtain gene expression counts. The number of samples contributed varies per individual and varies per tissue. As well, samples were collected at different developmental stages. The samples collected from 1 of the individuals are set aside for validation (referred to as the test set), leaving the samples collected from the remaining 25 individuals as the data to be used for model creation and model improvement (referred to as the training set). These datasets are provided in tabular form. Each entry represents the quantification of the expression level for the given gene in the given sample. Therefore, the set of genes analyzed is consistent across all samples. The training set contains 480 samples, each with expression levels for 46,430 genes. The test set contains 21 samples, each with expression levels for the same 46430 genes found in the training set.

The raw gene expression counts are first converted to transcripts per million (TPM). The steps of this preprocessing technique are as follows: 1) Divide the expression level by the length of the given gene in kilobases, 2) Divide by the summation of gene length normalized expression levels in the given sample, 3) Multiply by 1,000,000. The process of converting raw counts to TPM first normalized the expression levels by the gene length. More fragments are likely to map to longer genes and vice versa for shorter genes. These gene length normalized expression

levels are then adjusted for sequencing depth. Sequencing depth refers to the total counts attributed to a given sample. Given the inability to sequence all samples at the same time and with the same machinery, as well as the possibility that different protocols were used for different samples, sequencing depth normalization is an important step in being able to make comparisons across samples. Lastly, the expression levels are scaled by a large factor.

2.2 Data Preprocessing

TPM expression data is normalized by gene length and by sequencing depth, making it amenable for many downstream gene expression analyses. However, further preprocessing is often done before performing these analyses. Gene expression data is typically skewed with a relatively low number of very high values. As well, it is more biologically meaningful to look at proportional discrepancies as opposed to additive discrepancies. The expression data was $\log_2$ transformed with an offset in order to handle 0 values. The log-transformed data log-TPM is derived as follows for every x_{ij} entry in the $n \times p$ expression matrix X.

$$\text{transformed } x_{ij} = \log_2(x_{ij} + 1), i = 1, \ldots, n, j = 1, \ldots, p$$

Lowly expressed genes often do not provide a substantial amount of information and do not elucidate relationships between groups or genes. For example, removing lowly expressed genes when performing differential expression analysis between groups of samples will increase the power to detect significant differences given that the correction for multiple comparisons will be less stringent. Genes were deemed as lowly expressed if they met the following criteria: 1) No expression in at least 80% of samples, 2) The maximum expression across all samples was less than or equal to 2 TPM. 7733 genes were considered as lowly expressed and subsequently removed, leaving 38697 genes.

3 EXPLORATORY DATA ANALYSIS

3.1 Gene Expression Distributions

After data preprocessing, the data was first analyzed by looking at the mean expression levels across all genes. These distributions can be found in Figure 1. Mean expression levels from all individuals follow a similar distribution.

Picking the top g most expressed genes to serve as a feature set for predictive modeling may seem to be an advisable initial selection criterion. However, highly variable genes adjusted for the mean are likely to serve as better predictors. Figure 2 shows the relationship between the coefficient of variation (CV) and mean, as well as the distribution of CV values.

3.2 Dimensionality Reduction Visualizations

Dimensionality reduction techniques were used in order to quickly assess whether or not the data clustered based on certain sample metadata. Principal component analysis (PCA) was performed on the preprocessed expression data. As well, t-Distributed Stochastic Neighbor Embedding (tSNE) was performed on the preprocessed expression data. TSNE is a nonlinear dimensionality technique that is often used when analyzing single-cell data, but can be applied to any high-dimensional data. The main objective is to preserve the local structure of the data, while tSNE's main pitfall is its inability to preserve global structure and therefore its inability

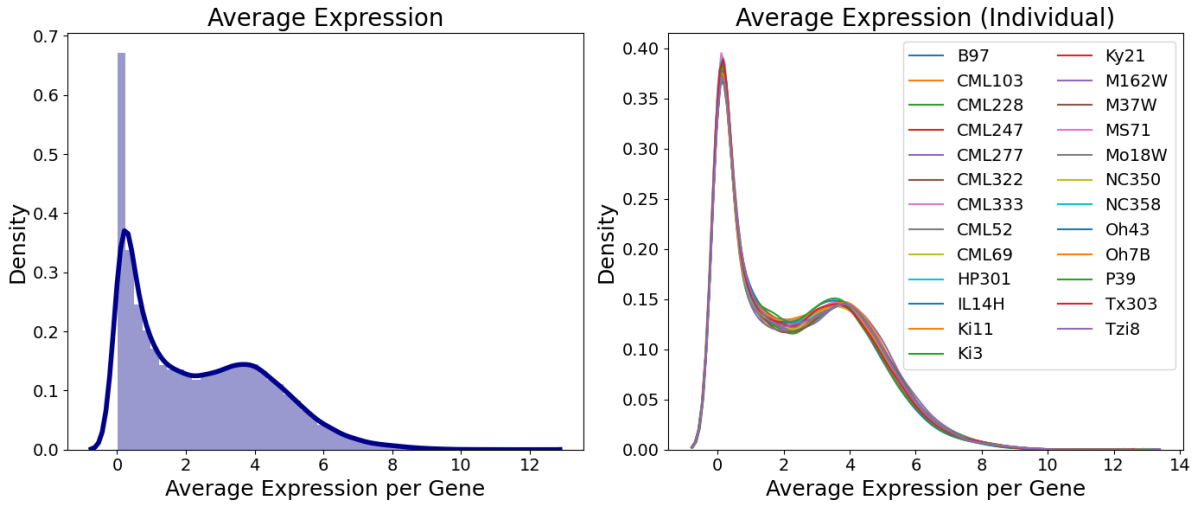

Figure 1: Distributions of the mean expression level per gene. The plot on the left shows the distribution when taking the mean expression level across all samples. The plot on the right shows 25 distributions derived from taking the mean expression level across samples from a particular individual.

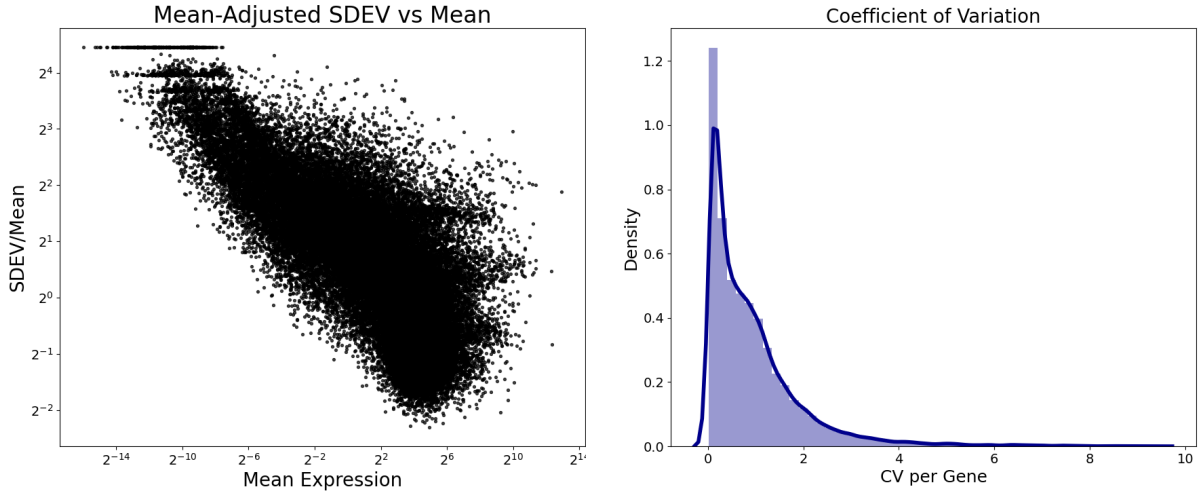

Figure 2: Left) Coefficient of variation vs mean for the expression level across all genes. Right) The distribution of the coefficient of variation values. The expression data used are unprocessed TPM values.

to draw conclusions from between-cluster distances. Explanations of these dimensionality reduction techniques can be found in the Appendix. Figures 3 and 4 show PCA and tSNE plots colored by individual and by tissue (organism part). Figure 4 indicates that there is distinct clustering by tissue. For this reason, it may be advisable to perform separate analyses and create separate predictive models per tissue.

•

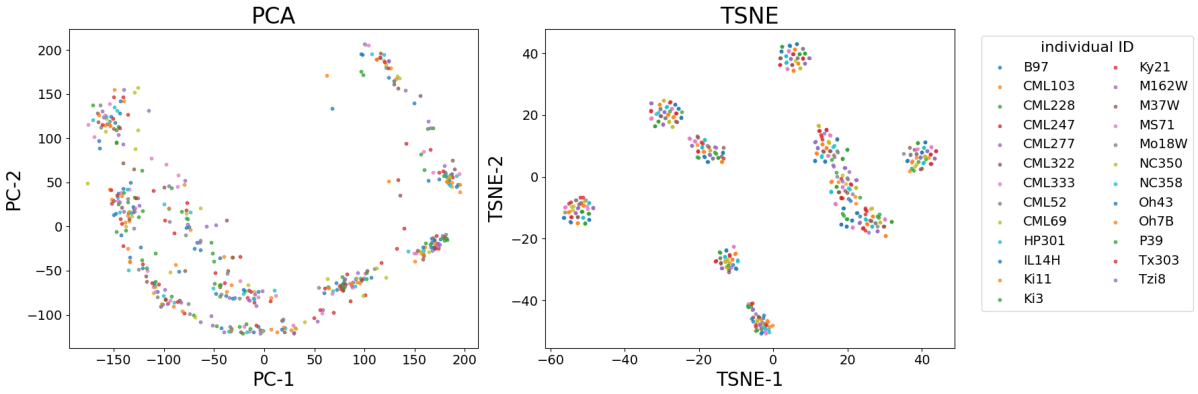

Figure 3: PCA and tSNE dimensionality reduction were used on preprocessed TPM expression values. The datapoints are colored by individual ID. There is no obvious clustering based on individual ID.

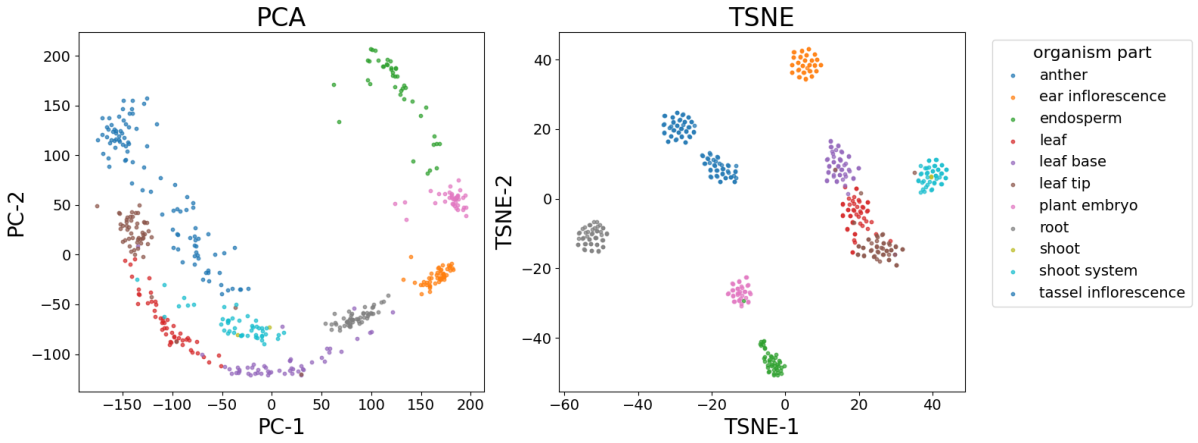

Figure 4: PCA and tSNE dimensionality reduction were used on preprocessed TPM expression values. The datapoints are colored by tissue. Distinct clustering based on tissue is present.

4 OVERVIEW OF METHODS

For Milestone 2, we approached our task on two separate routes. The first route further entails two main stages – feature selection and prediction. For each stage, we experimented with various methods that could later be mixed and matched.

Feature Selection Methods	Prediction Methods
GENIE3 Pairwise Correlation*	Various Regression Methods
Most Expressed 1,000 Genes^	
WGCNA Adjacency Matrix (Pairwise Correlation)*	
Landmark 1000 Genes^	

Table 1: Summary of feature selection and prediction methods
*Unique set for each target gene
^Common set for all the genes

Departing from the first route quite completely, we also approached the task from the angle of imputation.

4.1 Feature Selection Methods

For the following feature selection methods, our goal is to find a subset of genes, often 1,000 of them, that are most informative for predicting the expression values of each target gene.

4.1.1 GENIE3 Pairwise Correlation

GENIE3 is a tool aimed at recovering a gene regulatory network from expression data using tree-based ensemble methods. In other words, by running methods like Random Forests to infer on the expression values of each target gene, a list pairwise correlations can be derived. Then, for each target gene, we will select 1,000 genes that are most correlated with the target gene. The correlation values are not statistically significant but are indicators of the relative extents of pairwise correlations.

4.1.2 Most Expressed 1,000 Genes

As a first pass for our baseline model, we chose a representative subset of the genes by their Coefficient of Variation, or simply the standard deviation divided by the mean. This is a metric that aims to capture the genes with the most variability in their expressions, while normalizing over their average expression levels to compare between standard deviations of differing quantities.

The idea behind this approach is to avoid genes that are nearly constantly expressed (and hence would not help in any standard regression model), while at the same time ordering by genes that have relevantly high levels of expression such that their variability is not entirely anomalous or due to noise.

4.1.3 WGCNA Adjacency Matrix

Weighted Correlation Network Analysis (WGCNA) is a widely used tool to understand gene expression data, the relationships between genes, and the relationships between sample metadata and gene clusters. Given $p = 38697$ genes in the preprocessed expression dataset, a 38697×38697 adjacency matrix A, derived from a 38697×38697 correlation matrix, serves as the basis for WGCNA network construction. Each element in the adjacency matrix a_{ij} defines the relationship between nodes i and j. An intermediate value s_{ij} is calculated and denotes the strength of the relationship. s_{ij} for unsigned and signed networks are given below on the left and right respectively.

$$s_{ij} = |cor(\text{gene}_i, \text{gene}_j)| \qquad\qquad s_{ij} = cor(\text{gene}_i, \text{gene}_j)$$

Hard thresholding creates an unweighted network where $a_{ij} = 1$ given $s_{ij} \geq \tau$ and $a_{ij} = 0$ given $s_{ij} < \tau$ for the hard thresholding paramter τ. However, weighted networks created by using soft thresholding are better suited for modeling the continuous nature of biology and are therefore preferred. The adjacency matrix is defined as follows for a soft thresholding power β.

$$a_{ij} = s_{ij}^{\beta}, \beta \geq 1$$

WGCNA was performed on the data, generating an adjacency matrix. The optimal value for β is one that maintains a scale-free network topology. In other words, network connectivity that follows a power law distribution (i.e. relatively few nodes are highly connected and many nodes are lowly connected) is consistent with known biological interactions. A small proportion of proteins act as "master regulators" that interact with many different counterparts, activating or deactivating proteins, and upregulating or downregulating genes. β was determined to be 15. Until graph-based predictive models are implemented, assessing the validity of this network will not be possible.

4.1.4 Landmark 1000 Genes

We model our selection of the landmark 1000 (L1000) genes closely after the CMap project (Subramanian et al, 2017). The key idea is the assumption that there might be certain landmark genes that are of central importance to the genome and that might be good predictors of all the rest of the genes. The steps are:

1. Dimensionality reduction of the expression data using principal components analysis (PCA)
2. Cluster analysis to cluster the genes
3. "Bootstrapping" to obtain subsamples so as to obtain 100 different clusterings
4. Iterative peel-off to remove genes far from the centroids of the clusters
5. Consensus matrix to obtain the percentage of trials where each pair of genes are in the same cluster
6. Setting a threshold of the percentage, select the pairs of genes that are always in stable clusters and further group these pairs into clusters
7. By placing the processed clusters back in the reduced dimensions of the PCA and identify a single centroid, defined as the gene that has the least euclidean distances to every other gene in the processed cluster
8. The centroids of the processed clusters are pronounced as landmark genes
9. Repeat the process until 1,000 landmark genes are identified

4.2 Prediction Methods

As a first step, we decided to apply OLS "Ordinary Least Square Models" using the 1000 most expressed genes to predict each one of the next 1000 most expressed genes. The ranking metric used was the average expression level. After eliminating a total of 2,432 genes which were not expressed in any of the samples, we sorted the genes in two ways: 1) Using the average expression level per gene across all samples and 2) Using the coefficient of variation per gene across all samples. This allowed us to account for different scales across genes, ranking by variability in a more representative way.

In total we ran 1000 models per method using a train-test split of 70-30 (i.e., 70% for the training set and 30% for the test set). The models were fitted using the training set and R^2 of the test sets were obtained for further analysis of accuracy.

For sorting using the average expression level, the summary statistics for the R^2 for all models run are given in Table 2. As it can be seen, the results were really good in general.

Name	Value
Count	1000
Mean	0.747
Standard Deviation	0.285
Min	-2.521
25th Percentile	0.694
50th Percentile	0.829
75 Percentile	0.906
Max	0.992

Table 2: Summary Statistics for R^2. The average expression levels were used for sorting, determining the predictors, and determining the response variables.

In contrast, the results obtained when sorting by the Coefficient of Variation is given in Table 3.

Name	Value
Count	1000
Mean	-1.8995e+16
Standard Deviation	5.2427e+17
Min	-1.6541e+19
25th Percentile	-2.4124e+02
50th Percentile	-1.5933
75 Percentile	-1.4337e-01
Max	9.2419e-01

Table 3: Summary Statistics for R^2. The coefficients of variation for the expression levels were used for sorting, determining the predictors, and determining the response variables.

In general, almost all R^2 were close to 0, allowing us to conclude that the models were no better at predicting gene expression than naively predicting the average expression levels for the response genes.

Given the fact that using the same 1000 genes for predictions led to disappointing results, it was decided to use an alternative approach to get the features for predictions for each gene. GENIE3 allows to calculate pairwise correlations of genes by inferring gene regulatory networks from expression data. For each gene that we wanted to predict, we selected as features the ones that had 0.01 or higher correlation with it. For computational memory purposes we only used 4000 genes chosen arbitrarily for this part of the analysis.

Using a 70-30 percent train-test split, we ran a few baseline and ensemble models using both 1000 most expressive genes (sorted by Coefficient of Variation) to predict the next 1000 most expressive genes and with GENIE3 to select the features for the 4000 genes chosen for this analysis. The models that were ran were: a) Linear Regression, b) Lasso Regression, c) Lars Regression, d) Random Forest Regression, e) Ada Boosting Regression, f) Gradient Boosting (check the appendix for an explanation of the ensemble methods). Results of test R^2 are shown in table 5.

4.3 Imputation

Missing value imputation is widely used when analyzing gene expression data due to a phenomenon known as dropout. Either due to the detection threshold or inherent technical mishaps of the sequencer, the expression levels for certain genes for certain samples will not be registered. Instead of removing observations that contain missing values, often times missing value imputation is used in order to reconstruct these expression levels from the available information (i.e. the non-missing entries). Given the abundant documentation on the performance of imputation methods in the context of gene expression data, several imputation algorithms were re-purposed for predictive modeling.

Mean, k-nearest neighbors (kNN), and variational autoencoder (VAE) imputation models were built and compared. Each model was trained on the training set consisting of samples from 25 individual maize plants (n = 480). The training set did not contain any missing values. Trained models were used to impute missing values in the test set with samples from 1 individual maize plant (n = 21). The imputation models are described as follows.

4.3.1 Naive Mean Imputation

The naive mean imputation model predicts the mean expression level for a given gene. First, the mean expression level for every gene in the full training set is calculated. The mean expression levels are used to impute missing values in the test set for those genes that contain missing values. A gene's missing value(s) are imputed with the mean expression level of the same gene derived from the training set.

4.3.2 k-Nearest Neighbors Imputation

The kNN imputation model constructs a kNN graph by deriving the k nearest neighbors for every gene. The nearest neighbors are defined as those genes among the k^{th} closest genes based on the euclidean distance. The distance measure $d(p, q)$ between genes p and q is defined as follows.

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (x_{ip} - x_{iq})^2}$$

A missing value x_{ip} found in the test set is imputed with the average expression level of the k nearest genes at the given observation.

$$x_{ip} = \frac{1}{k} \sum_{j=1}^{k} x_{ij}$$

4.3.3 Variational Autoencoder Imputation

Autoencoders are a family of neural networks with a 2-part architecture comprised of an encoder and a decoder. The encoder takes an input X and transforms the input into a latent representation Z. The decoder takes as input Z and outputs a reconstructed version of the input $\hat{X}$. The neural network aims to minimize the reconstruction error or reconstruction loss

$L(X, \hat{X})$ that quantifies the discrepancy between the input and the output. In most cases, the Z is constrained to a lower-dimensional space as compared to X. Variational autoencoders, instead of using an encoder to map observations of the input to points in a latent space, use an encoder to map observations of the input to their respective latent distributions. In practice, the latent distribution corresponding to each observation is normal. The decoder samples from the latent distributions before reconstructing the input. The sampling step effectively regularizes the autoencoder, making the model less likely to overfit to training data. Variational autoencoders have been shown to perform well for imputing missing values in gene expression data (**ADD REFERENCE**).

4.3.4 Imputation Performance

The 3 imputation models were tested under two distinct scenarios. In scenario 1, a specified proportion of the entries in the expression matrix are selected and subsequently masked. In scenario 2, a specified subset of genes are selected and all entries corresponding to the selected subset of genes are subsequently masked. Scenario 2 is analogous to the case where a subset of genes are used as predictors in order to determine the expression levels for the response genes for which no data exists. Representations for scenario 1 and scenario 2 are given in Figure 5. Each model was tested under varying levels of missing data severity. For both scenarios, 5%, 10%, and 50% of the entries in the test set were masked, these entries posing as missing values. For every scenario-missing data severity pair, missing values were randomly generated 5 times. The average R^2 values are reported in Table 5. The computation for the R^2 values is based on the true values and imputed values for the missing data and does not take into account all entries in the test set as this would lead to an artificial inflation of the given model's performance.

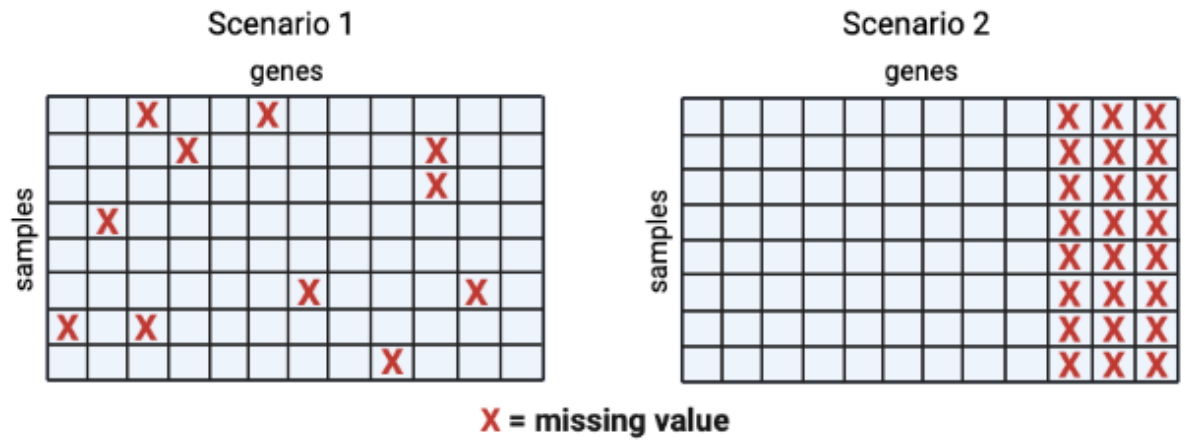

Figure 5: Scenario 1: A proportion of entries are randomly selected and masked. Scenario 2: A subset of genes are randomly selected and all entries corresponding to the selected genes are masked.

—	Mean Imputation R^2	kNN Imputation R^2	VAE Imputation R^2
scenario 1: 5% missing value rate	0.7231	0.9073	0.8985
scenario 1: 10% missing value rate	0.7235	0.9075	0.9052
scenario 1: 50% missing value rate	0.7223	0.8801	0.8936
scenario 2: 5% missing value rate	0.7135	0.9070	0.9034
scenario 2: 10% missing value rate	0.7206	0.9068	0.8981
scenario 2: 50% missing value rate	0.7217	0.8923	0.8941

Table 4: The average R^2 values for several imputation methods applied to different scenarios of missing values.

5 RESULTS

5.1 Feature Selection & Prediction Results

5.1.1 GENIE3 for Feature Selection Plus Various Prediction Methods

As the reader can see in table 5, if we use the same features (i.e., the 1000 most expressive genes by Coefficient of Variation) for all predictions, we get disappointing results of the test R^2 for all models. On the other hand, If we use GENIE3 to select the features per gene, we get significantly better results. As expected, the ensemble models resulted to be the better models. Random Forest achieved the highest mean R^2 with 0.6483. 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. For each gene that we have to predict, we have to calculate the pairwise correlation with all of the other genes to select the features that we're going to use for the respective regression.

Given these results, we decided to pursue several more feasible feature selection methods and models that will likely provide better performance and which are going to be discussed in the next sections of this report.

—	Mean of R^2 GENIE3	Mean of R^2 1000 Most Expressive Genes
Linear Regression	0.4895	-1.899e+16
Lasso Regression	0.5061	-27.0659
Lars Regression	0.4889	-7.8803
Random Forest Regression	0.6483	-1.6092
Ada Boosting Regression	0.5755	-2.3085
Gradient Boosting	0.6285	-4.0844

Table 5: Average R^2 values for regression models using both feature selection methods

5.1.2 L1000 for Feature Selection

In this section, we attempt to recreate the L1000 landmark gene (Subramanian et al., 2017) approach that uses aggressive clustering techniques to select a representative gene set on which to predict the rest. In particular, this approach emphasizes the need to find a highly informative subset of the gene expression data, that ideally captures the variation allowing us

to predict the expression of the held-out majority group. The main benefits of this approach lie in the fact that of the existing literature, this is one of the few regression-oriented approaches that have been attempted with purely gene expression data, and that they seemed to have decent accuracy, measured by the ability of the landmark 1000 to recover 81-83% (paper cites varying numbers) of the missing gene expressions.

The main challenges are that their experiment was conducted with many orders of magnitude more effective data, in both sheer quantity (10,000+ individual samples) and data variation (164 introduced perturbations in cells), as well as inference targets (10,000 genes studied instead of 40,000). In this section, we attempt to replicate their methods and determine the efficacy of the landmark gene approach at our data scale.

5.1.3 L1000 Methodology

The provided code in Subramanian et al.'s (2017) paper and respective GitHub repositories were extremely difficult to navigate and reproduce, spanning multiple platforms (Matlab, Python, Java, R) where critical analytical steps had unique implementations *across* different platforms, meaning that there was no complete Python pipeline (requiring analysis to be partly done on the other platforms). On top of this, the code when studied in detail revealed that they made huge assumptions about how the data was structured, tailored specifically to their dataset. Finally, the actual step of selecting the L1000 landmark genes and the processing required to do so was elided from their pipelines, where downstream processing referred to a hardcoded "L1000" data file instead of upstream pipeline components to do this processing.

In light of these challenges, we reproduced the critical elements of their pipeline and broke down the necessary steps into reproducible components using simple, existing library implementations such as clustering tools from Scikit Learn. The basic selection algorithm is as follows:

1. **Dimensionality reduction:**
   In order to perform Euclidean-based clustering, the original data dimensions are transformed into a lower dimensional representation. Here, like in the referred paper, we chose Principal Component Analysis and decomposed each gene into the top 50 representative principal components. We visualize the PCA decomposition using a further tSNE dimensionality reduction in Figure 6.

2. **K-Means Clustering:**
   We then cluster on the first 50 principal components, using a simple K-Means clustering. In order to speed up computation over the width of the dataset, we use Mini Batch K-Means clustering. Figure 7 is a visualization of one sample of K-Means clustering, overlaid over the tSNE representation of the PCA components.

3. **Consensus Matrix Construction**
   The next step involves repeating the K-Means sampling on "bootstrapped" samples, each representing a randomly sampled 75% of the total gene count. As done in the paper, this repeated K-Means sampling will be repeated 100 times. Across all 100 clustering runs, a consensus matrix is then constructed, where the pairwise ratio of the number of times two genes end up clustered together across the 100 runs is represented.

   The consensus matrix is then thresholded for clusters above a certain ratio, arbitrarily .8 for the original approach. Clusters that remain in the consensus matrix with genes that are highly clustered together are then the targets for further processing.

Figure 6: PCA decomposition of all corn genes, visualized using further tSNE dimensionality reduction. Some small clusters of similar genes, hinting at gene regulatory subnetworks, can be seen.

One of the primary challenges this step revealed was the lack of variation in our source data set. The greatest source of variability in our gene expression levels comes from the 10 different tissue samples. However, there are no further perturbations introduced per sample, which in comparison to the 164 experimentally induced perturbations in the original paper causes a number of issues like large cluster size.

In Figure 8, we see the average of cluster sizes across all 100 K-Means cluster attempts. The variability was low, and each individual cluster attempt demonstrated the visualized behavior of very large clusters at the top-end with an average of 4000 members. Although this is the largest such cluster and the next clusters range in the hundreds, the ideal cluster size for a data set of 40,000 genes and 1000 landmark genes is at least an order of magnitude smaller (roughly 40000/1000 = 40).

4. **Iterative Peel-Off**
   In the next step of the pipeline, the top clusters identified are utilized by selecting the single gene closest to their center. The center is defined as the centroid of the cluster in PCA space, with a sample visualization in Figure 9.

   The iterative peel-off methodology then removes all other cluster genes from consideration, and only keeps the gene closest to the centroid as a landmark gene. Once a few landmark genes are identified, the entire process of 100 K-Means clustering attempts is

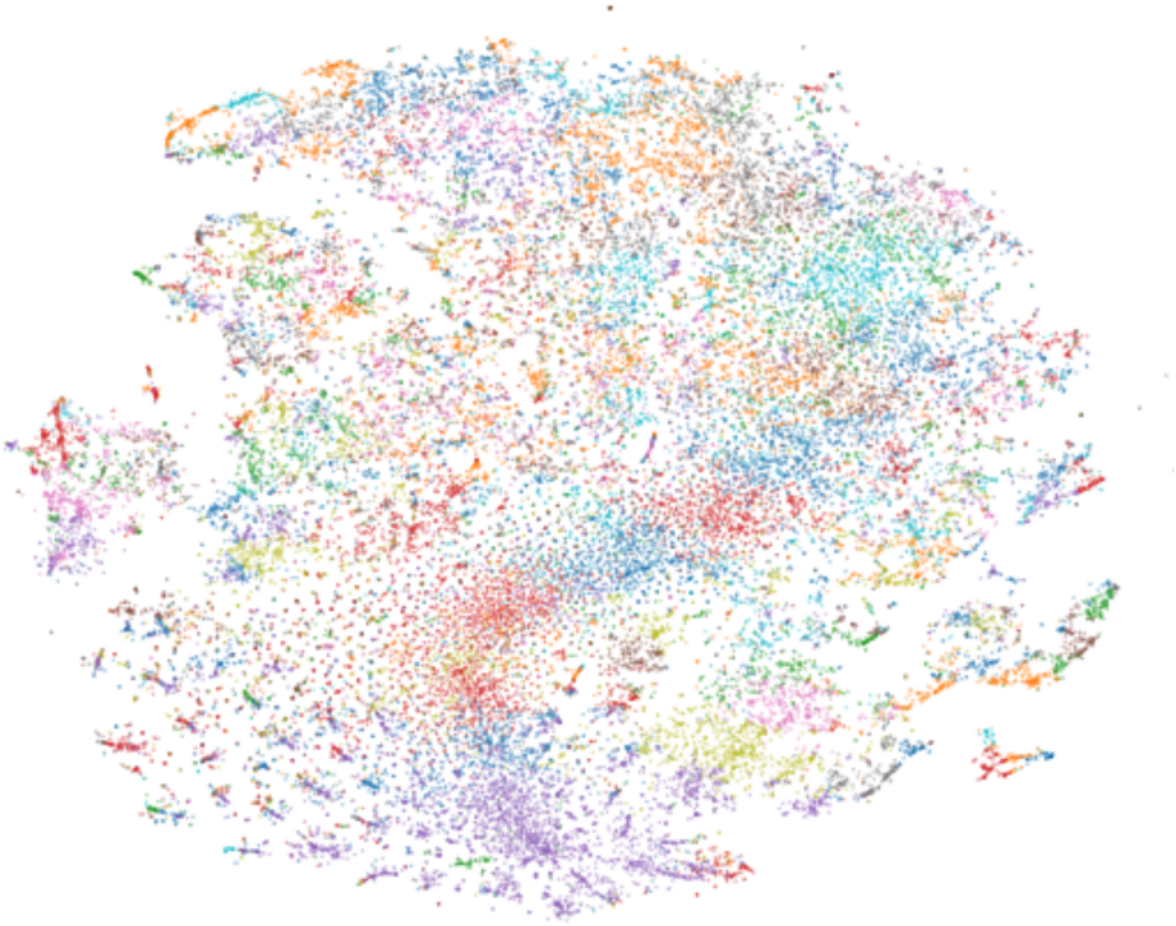

Figure 7: K-Means clustering applied to the first 50 principal components, again visualized by tSNE. We chose $k = 100$ for each iteration of the iterative peel-off process described below. While cluster overlap between tSNE and and K-Means projections seem fairly weak visually, many of the larger clusters (note purple cluster at bottom) were particularly stable across multiple K-Means iterations.

repeated, and another consensus matrix is created. The hope is that without the genes of the previously identified clusters for landmark selection, a new highly related cluster will surface as candidates for the next few landmark genes.

In practice, this approach on our problem ran into a number of significant challenges. The first of which is already explored above, and primarily stems from the lack of data variability to form tight enough clusters such that we do not remove all of our data before we get a chance to form a large enough representative set. This is ameliorated by the random iterative sampling, and in truth with tight thresholds (pairs of genes show up in >0.8 of clustering attempts) we can see small clusters that might suggest appropriate landmark genes.

The main issue however, is one purely of computation. We have an exponential data scaling issue, even more so than the original paper's 10,000 human gene set under analysis. Consider a single consensus matrix shaped 40000×40000: if we use standard floating point values of 8 bytes, we are looking at $40000 \times 40000 \times 8 = 25600000000$, or 25.6 GB of data for a single instantiation. This can be reduced with lower precision representations depending on what we need, but the sheer computation of multiple arrays requires ma-

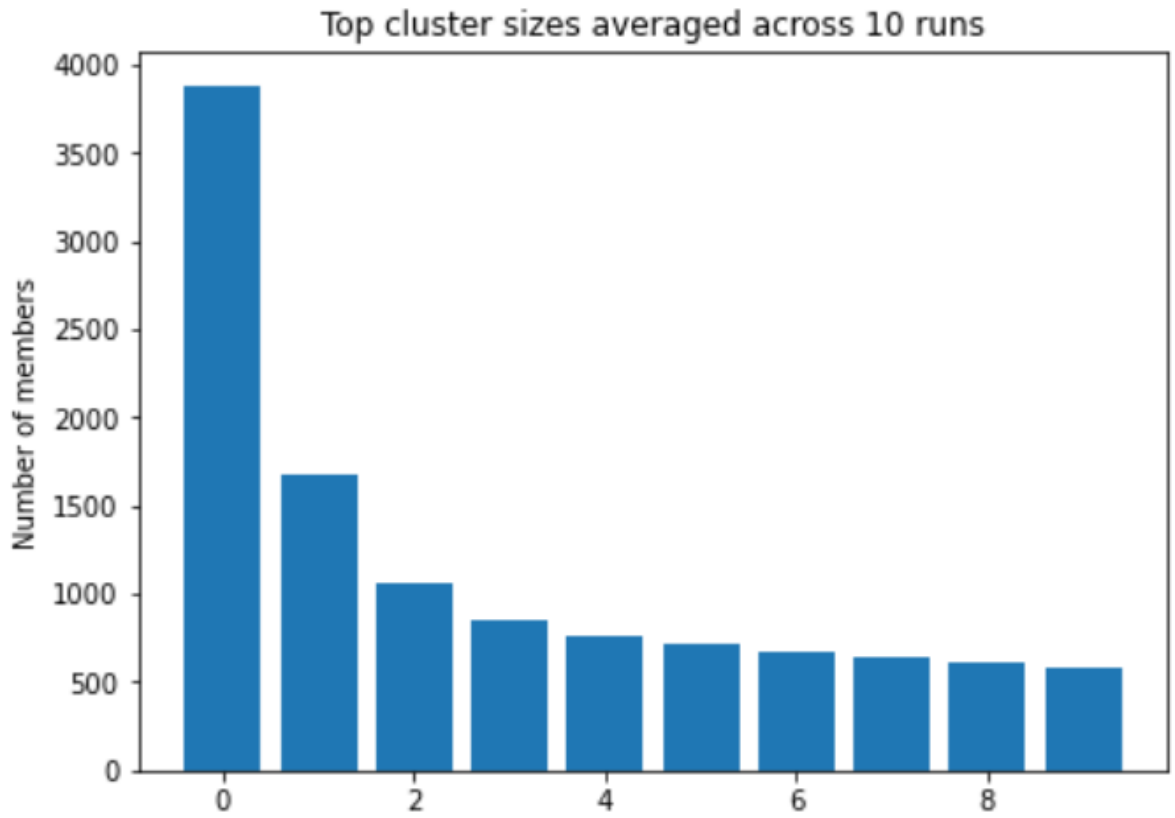

Figure 8: A chart of the largest cluster sizes, averaged across 100 iterations of K-Means with k=100. The largest clusters are disproportionately large, on average around 4000 gene members, easily eclipsing the target of 1000 genes as a representative sample for the entire set. These clustering averages remain relatively consistent even with k=1000, as the granularity is developed at the tail end.

chines with larger than conventional memory sets. Note that at 10,000 genes, these sizes are manageable on standard 16GB machines. We are currently considering methods of scaling our computation and working on cloud computing with larger memory resources.

6 NEXT STEPS

6.1 Problem Analysis

One of the central motivations behind our problem is the attempt to understand the full relationship graph of all expressed genes, or more specifically, reconstructing some kind of Gene Regulatory Network. GRN's in the literature have been well studied for more than a decade, but the area remains largely an "unsolved problem" (Saint-Antoine et al., 2019) and is complicated by a bevy of computational challenges.

More formally, work on GRNs fully encompasses our current problem setting as follows: we consider N genes and represent their expression levels (the entirety of the data we have) as random variables $\{X_1, X_2, ..., X_N\}$. Each X_i thus represents a node in our network, and any relation $r(X_i, X_j)$ describes an edge where r models the regulatory relationship between the two genes. These relations can be modeled as directed, implying causality, or with signs, hence

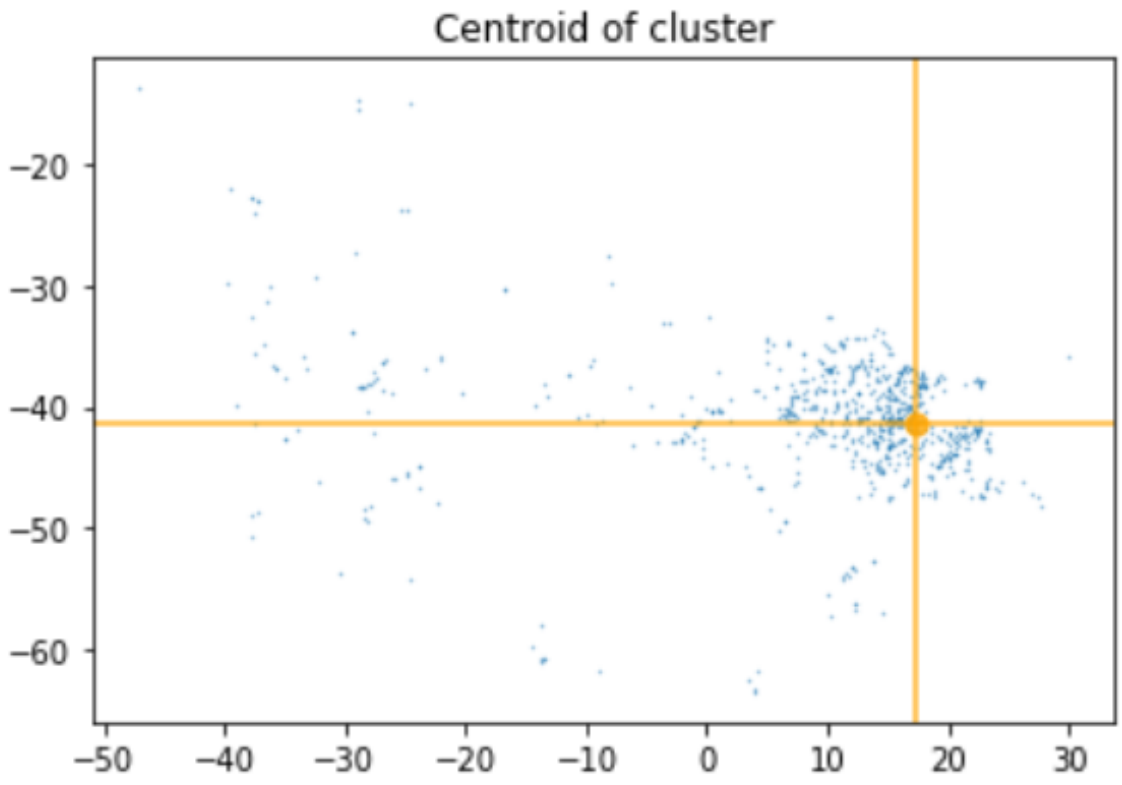

Figure 9: A sample cluster centroid, of which we then choose the single landmark gene closest in Euclidean distance to this centroid. As part of the iterative peel-off process, all other genes in this cluster will be removed except for the chosen landmark gene.

representing downregulation or upregulation.

This formal description of the problem leads to the natural formation of any model that wishes to describe a GRN as some subset of all possible relations $R = \{r \mid r : (X_i, X_j) \mapsto s\}$, where s is typically some score representing the importance of that relation. Indeed, the vast majority of methods in reconstructing GRNs use precisely such a metric, and one of the most popular competitions Dream Challenges ([homepage - Dream Challenges](http://dreamchallenges.org/)) provides gold-standard evaluation data where the target labels are exactly these edge importance values, as verified experimentally or through other modalities.

One key discrepancy between our current problem formulation (i.e. regression on subset of most highly expressed to find expression levels of the rest) and the existing literature is that our evaluation metric is directly based on expression levels, while all other existing methods try to evaluate based on real biological understanding of gene regulatory pathways. There are a number of clear downsides to metrics based purely on expression levels:

- Consider genes with low variability (almost always expressed at a constant level, or almost always 0, etc.). All aggregate metrics of accuracy over any test set Y would be arbitrarily inflated in accuracy by inclusion of any low-variability genes as all regression methods would simply set feature coefficients to 0 and be a constant predictor.
- Even with a reasonably accurate regressor, it is not trivial to then extrapolate that information and establish which genes are related, what sub-networks exist, what kind of

motifs of regulatory action are revealed, etc. Physiologically speaking, we have no experimental variation in our data, and cannot expect to capture consistent patterns tested by perturbation, and rely solely on random individual variation.

- As follows, it will also be hard to generalize the utility of results based on cross-validation of these metrics, given the two-fold difficulty of lacking experimental variation as well as any basis of interpretability.

6.2 Existing Approaches

There are a number of classes of approaches in reconstructing Gene Regulatory Networks, each of which begin with the same kind of data that we have (purely gene expression levels, without other modalities of data). Each of these are potentially interesting avenues to pursue, and are worth further experimentation and consideration in our attempt to derive genuinely interesting insight from our data. Figure 5 from Saint-Antoine et al. (2019) describes each class with their respective qualities and advantages/disadvantages.

Algorithm Class	Temporal Data Required?	Directionality	Advantages	Disadvantages	Examples
Correlation	No	Undirected	• Fast, scalable • Detection of feed-forward loops, fan-ins, and fan-outs	• Possibly over-simplistic • False positives for cascades	WGCNA [13] PGCNA [14]
Regression	No	Directed	• Good overall accuracy	• Bad detection of feed-forward loops, fan-ins, and fan-outs	TIGRESS [15], GENIE3 [16], bLARS [17]
Bayesian - Simple	No	Directed	• Performance on small networks	• Performance on large networks. • Inability to detect cycles	[19,20]
Bayesian - Dynamic	Yes	Directed	• Performance on small networks • Detection of cycles and self-edges	• Performance on large networks.	[21]
Information Theory	No	Undirected (at least in simplest form)	• Detection of feed-forward loops, fan-ins, and fan-outs • Similar to correlation methods, with better accuracy	• False positives for cascades	ARACNE [25], CLR [26], MRNET [27], PIDC [28]
Phixer	No	Directed	• Parsimonious output due to pruning step.	• Possible loss of overal accuracy due to pruning step (this can be removed if the user chooses)	[31]

Figure 10: The classes of models used throughout literature in attempting to reconstruct Gene Regulatory Networks.

- **Correlation:** This class of network inference works purely off correlation matrices calculated from the data, and uses varying threshold strategies to try to establish network edges. Baseline models involving WGCNA fall into this category, though their use in predictive settings will devolve equivalently into basic linear regression.

- **Regression:** The setting that most resembles our problem statement. Models in this class like TIGRESS and GENIE3 all work the same way by predicting the expression of a leave-one-out gene from the rest of the expression levels, using some kind of regressor (Least Angle Regression for TIGRESS, and random forest for GENIE3). This is a potential approach for the most direct next step of our modeling.
- **Bayesian:** Either classified simple or dynamic, this class of analysis is computationally expensive and thus well suited for smaller sized networks, but perform quite well in the literature. In the case of dynamic Bayesian networks that incorporate temporal or pseudo-temporal features, self-regulating genes (modeled by self-edges) can also be discovered.
- **Information theory:** Similar most to correlation approach, these models use mutual information or Shannon information to understand distributional characteristics and test the importance of gene relations under this framework. A popular approach that can deal with multi-step relations, unlike regression approaches.

6.3 Graph Neural Networks

We have done extensive research into many classes of potential Graph Neural Networks to use, and will look into compatibility for the next milestone: StellarGraph, PyTorch Geometric, Deep Graph Library to name a few complete code libraries that exist with good candidate selections. The primary difficulty of usage based on current experimentation is formation of the data and in particular evaluation strategy: prediction on single-attribute nodes is not a very common task in generic graph neural networks currently, and will require an effort of coadapting the models.

To be more specific, graph neural networks aim to solve the following classes of problems:

- **Node classification:** where we infer the class of nodes, e.g. tissue type. We can adapt this to a regression setting, but given that each node's data is essentially one-dimensional (just expression value), this has not been the easiest fit as many networks in this class expect wide nodes with a fair amount of information in each.
- **Link prediction:** inference on missing edges and the relationships of nodes. This is much more prevalent in work in research and pertinent for describing GRNs, but our predictive setting cannot leverage this class of models unless we redefine objectives.
- **Community detection:** this would be an interesting avenue to pursue to do unsupervised clustering of related genes, but does not directly contribute to predictive efforts. Worth keeping on the horizon as potentially a preprocessing step.
- **Graph classification:** classification on the level of an entire graph (i.e. some phenotype given full expression), less relevant as we have no labeled data to this effect.

Combining the power of graph neural networks with the existing approaches in literature in reconstructing GRNs is one of our immediate future directions, and a potential area for novel research and genuine contribution given that we can resolve formulations of evaluation metrics in a way that can adapt to network architectures.

7 APPENDIX

The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of a large number of interrelated variables while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables.

Let X be the matrix of all the independent variables (in this case, the genes) where each column is a gene and each row corresponds to a sample of maize. We then calculate the mean for each column and subtract the respective mean from each variable. This gives us a matrix whose columns has zero mean. We then divide each data with the standard deviation of each columns to center the matrix. We call this centered matrix Z. Covariance matrix of Z is then calculated by multiplying it with Z^T, which is the transpose of matrix Z.

After obtaining the covariance matrix of the centered matrix Z, we then find its eigenvalues and eigenvectors. Since the covariance matrix is a semi-definite matrix, the eigenvalues and eigenvectors can be calculated by eigen value decomposition. The eigen decomposition of $Z^T Z$ is where we decompose $Z^T Z$ into PDP^{-1}, where P is the matrix of eigenvectors and D is the diagonal matrix with eigenvalues on the diagonal and values of zero everywhere else. The eigenvalues on the diagonal of D will be associated with the corresponding column in P — that is, the first element of D is λ_1 and the corresponding eigenvector is the first column of P. This holds for all elements in D and their corresponding eigenvectors in P. We will always be able to calculate PDP^-1 in this fashion. Take the eigenvalues $\lambda_1, \lambda_2, \ldots \lambda_n$ and sort them from largest to smallest. In doing so, sort the eigenvectors in P accordingly. We call this sorted matrix of eigenvectors P^*. (The columns of P* should be the same as the columns of P, but perhaps in a different order.) Note that these eigenvectors are independent of one another.

We then calculate the matrix with principal components $Z^* = ZP^*$. This new matrix, Z^*, is a centered/standardized version of X but now each observation is a combination of the original variables, where the weights are determined by the eigenvector. Since columns of P^* are independent, the columns for Z^* are also independent.

One downside from the PCA is that it might lead to poor visualization especially when dealing with non-linear manifold structures. Instead of being concerned with preserving large pairwise distances to maximize variance as the PCA, the TSNE focuses on preserving only small pairwise distances or local similarities. It calculates similarity measure between pairs of instances in the high dimensional space and in the low dimensional space, and uses a cost function to optimize between these measures.

Several ensemble methods used for the regressions were Bootstrap Aggregating through Random Forest and Boosting. Bootstrap Aggregating is an ensemble technique in which we build many independent predictors/models/learners and combine them using model averaging techniques (e.g., weighted average, majority vote or normal average). The essence is to select N bootstrap samples, fit a classifier on each of these samples, and train the models in parallel. The results of all classifiers/regression are then averaged into a Bootstrap Aggregating classifier/regression. One example of a Bootstrap Aggregating model is Random Forest, where decision trees are trained in parallel.

On the other hand, Boosting allow us to train models sequentially, instead of modelling them parallelly. Each model focuses on where the previous classifier performed poorly.It makes N decision trees during the training of data. As the first decision tree/model is made, the records

incorrectly classified are given more priority. Only these records are sent as input for the second model. The process will go on until the number of base learners we specified initially . One important thing to note is that the repetition of records is allowed with all boosting techniques, this guarantees independence across them.

The two Boosting techniques used were: 1) Adaptative Boosting and 2) Gradient Boosting. Adaptive Boosting (AdaBoost) uses multiple weak models to build a strong one. The base model could be any from Decision Trees (often the default) to Logistic Regression, etc.

When decision trees are used, instead of using trees with no fixed depth such as in random forest, Adaboost only uses nodes with two leaves, Decision Stumps (decision trees with a single split). These are the weak learners in AdaBoost. They work by putting more weight on difficult to classify instances and less on those already handled well. The weights are re-assigned to each instance, with higher weights to incorrectly classified instances.

Gradient boosting relies on the idea to repetitively leverage the patterns of residuals and strengthen a model with weak predictions and make it better. Once we reach a stage that residuals do not have any pattern that could be modeled, we can stop modeling them (otherwise it might lead to overfitting); we use residuals as target variables in the sequential models. We' re minimizing our loss function, such that test loss reach its minima.

Saint-Antoine, M. M., Singh, A. (2020). Network inference in systems biology: recent developments, challenges, and applications. Current Opinion in Biotechnology, 63, 89–98. https://doi.org/10.1016/j.cop

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., Gould, J., Davis, J. F., Tubelli, A. A., Asiedu, J. K., Lahr, D. L., Hirschman, J. E., Liu, Z., Donahue, M., Julian, B., Khan, M., Wadden, D., Smith, I. C., Lam, D., Liberzon, A., ... Golub, T. R. (2017). A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell, 171(6), 1437–1452.e17. https://doi.org/10.1016/j.cell.2017.10.049