MAJOR TO DO:

**Reviewer 1 -**

1. Extended comparison to alternative methods:

- Compare to kNN, factor analysis (or N/A?).

- Prediction accuracy vs. gene abundance.

2. Robustness wrt annotation:

- Impact of false positive gene sets and/or missing markers?

- Impact of program size, variance in training population, mean expression?

- Prediction accuracy of single genes vs. number of programs they’re in?

3. Algorithmic details:

- Run time?

- Measure of confidence?

4. Limitations in datasets with single-gene perturbations:

- Can we detect single/small number of gene changes?

**Reviewer 3 -**

1. Rephrase, “reconstruct”

2. Accuracy vs. #-samples in training.

**Reviewer 4 -**

1. Robustness analysis of pathways like reviewer 1. Random pathway performance.

2. Report other measures of reconstruction performance. e.g. RMSE.

3. Clarify usage of intra-submission accuracy.

4. How good is the measured gene as a proxy for program expression? See reviewer 1 baselines.

**Reviewer 5 -**

None major.