1. General Steps:
   1. Preprocessing
      1. input: expression matrix
         1. row: cell
         2. column: genes
      2. filter:
         1. num\_genes <200 and > 5000, discard (using histogram). Outlier detection algorithms could be applied here
         2. filter out mitochondrial gene transcripts gene
         3. discard low expression genes, usually fewer than 3 cells in the data. Why? To be researched
      3. Inference:
         1. To infer the interaction between TF and target genes
            1. For each target gene, build a ensemble model
            2. How to measure the importance for each TFs?

Use feature importance based on the information gain

https://github.com/vahuynh/GENIE3/blob/master/GENIE3\_python/GENIE3.py

For each tree, we use the decision\_tree.computa\_feature importance function to compute the feature importance for each tree and use the mean for all the trees

Issues: this prefer to high cardinality categorical features and numerical features, might not be useful to generalize( see: <https://stats.stackexchange.com/questions/450703/is-feature-importance-in-random-forest-useless>)

Random forest is non-deterministic model

SHAP is more appropriate?

* + - 1. Module generation
         1. Get top-50 target response gene for each TF
      2. TF-regulon prediction
         1. Hidden-Markov,