1. General Steps:
    a) Preprocessing
        i. input: expression matrix
            1. row: cell
            2. column: genes
        ii. 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
        iii. Inference:
            1. To infer the interaction between TF and target genes
                a) For each target gene, build a ensemble model
                b) How to measure the importance for each TFs?
                    i. Use feature importance based on the information gain
                    ii. https://github.com/vahuynh/GENIE3/blob/master/GENIE3_python/GENIE3.py
                    iii. 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
                        1. 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)
                        2. Random forest is non-deterministic model
                        3. SHAP is more appropriate?
            2. Module generation
                a) Get top-50 target response gene for each TF
            3. TF-regulon prediction
                a) Hidden-Markov,