Reflections

After reading Task 3, I realized that this is a question that gives a lot of freedom while covering a large number of cases. Cases that I know little about, due to my higher education which is unrelated to biology. However, I decided to send my submission to share my, very simple insights, but maybe it would be helpful.

Idea

I noticed that during task 2, at least 2 metrics were introduced, for different application. During this task, my goal is to find uniwersal solution for all cases, where the goal is to score cell type distribution, basing on predefined distribution pattern Q. According to Keep It Simple rule, in such general cases I would use general solution and I think that $EuclideanDistance$ fits here perfectly. So define Q as the desired cell state proportion vector, Q_i as cell state proportion vector after knockout gene i, and $EuclideanDistance(Q,\ Q_i)$ as $\sqrt{\sum_{n=0}^{4}(q_n\ -\ q_{in})^2}$ where n is dimension index. We would like to have a score (the higher, the better), not a distance, so we can propose score $Score_{qqi}\ =\ \frac{\sqrt{2}\ -\ EuclideanDistance(Q,\ Q_i)}{\sqrt{2}}$. $\sqrt{2}$ is maximum $EuclideanDistance(Q,\ Q_i)$ (sum of all 5 dimensions of Q or Q_i is equal to 1) and it is used for changing sign (now we have score: the higher, the better) and normalization ($0\ \leq\ Score_{qqi}\ \leq\ 1$). If $EuclideanDistance(Q,\ Q_i)\ =\ 0$ then whole score is equal to 1 and it is lower otherwise. $EuclideanDistance$ comparing to $L1$ distance prefers an error spread between dimensions rather than cumulative in one dimension, where $L1$ treats both cases the same, which in my opinion is worse.

Estimation Q_i

In my previous tasks, I estimate Q_i basing on P_i, the gene expression distribution of the cells obtained by knocking out gene i. To do this, I trained classifier (RandomForest but it could be any classifier e.g. K-NN) on given dataset, using cell type as a label and expression as an input. Then I was able to assign each sample from distribution P_i a label. After I was able to compute Q_i normalizing counts of states by sum of all samples.
I this experiment I assumed that we have samples for P_i distribution, but having joint probability distribution we are able to generate them.