**Objective Function for Scoring Gene Perturbation**

Using the predicted cell state proportions for each gene generated by the trained model in Challenge 1, we can rank gene perturbations in terms of their capacity to move cells from one state to another with respect to a desired *Q*. Aside from ranking genes based on their distance to the desired *Q* as was done in Challenge 2A and 2B¸ a more holistic objective function can be created that takes into account several variables that influence the effectiveness of a gene perturbation in changing cell states for cancer immunotherapy. In addition to *Q*¸ we incorporate the following variables: change in gene expression after perturbation, growth rate *G*, and a statistic *s(*) that represents a transformed summary the gene expression of that takes into account classification boundaries.

Given the gene expression distribution of unperturbed cells of 15,077 x *n* dimension, where *n* represents the number of cell samples, the states of each cellcan be obtained by using an unsupervised machine learning algorithm which clusters cells with similar gene expression profile. This was demonstrated in the UMAP provided at the start of the challenge from which cells were assigned a state classification. Thus, from , we use the cell state proportion of the unperturbed cells as a baseline to compare the predicted from a gene perturbation. We want to include a variable that reflects the ability of the perturbation to change the gene expression profile of T cells. A perturbation that barely induces a change in cell state proportion with respect to the baseline unperturbed cell state vector indicates that the gene perturbed has no importance in influencing gene expression from baseline. Let be the cell state proportion of the unperturbed cells obtained from screening . We represent the change from unperturbed to perturbed cell state as follows:

= =

Where a score of 0 indicates no change cell state proportions and a score of 1 means complete empirical change. Variables with *i* subscriptrefers to the state of the perturbed cell, whereas variables with o subscriptrefers to the state of the unperturbed cell.

Similarly, we compare the distance between the perturbed cell state proportion and the desired *Q*. The distance formula is inversed such the smaller the difference where the denominator approaches 0, the larger the obtained value. By doing this, we obtain information if the predicted perturbed state is similar to the desired *Q* to rank the genes. We represent the difference between the perturbed cell state and the *Q* desired as follows:

= =

Where a score closer to 1 means high similarity to the desired *Q.*

Moreover, growth rate *G* is reflected by the cycling state. Cycling state of a T cell indicates their tendency to undergo self-renewal and cell division to proliferate. Calculating the Spearman correlation between number of cells and cycling state from the training dataset, we observe a direct relationship between the two variables, with a correlation score of 0.47 and p-value of 4.81e-5. A low cycling state probability indicates a low growth rate.

Finally, the statistic takes the predicted perturbation cell state and transforms it to a new variable such that uncertainty in cell state classification is penalized. By doing this, a more stringent classification of a cell is preferred between tumor-killing cell states and a weak cell state. Here, we consider *a* and *b*, referring to progenitor and effector states, respectively, as active tumor-killing cell states, and *c* as the weak cell state. As seen in the equation below, we penalize similarity in values by to subtracting the absolute difference between the variables from 1 and add it as a penalty term in the objective function. Through this, we give higher ranking to cells with more strong probability of being in an active state than being in equal probability with a weak state. For instance, a [0.4, 0, 0.4, 0.2, 0] should be penalized as the cell has equal probability of being in active and weak cell state which gives uncertainty to its classification.

Where *a*, *b,* and *c* are the progenitor, effector, and terminal cell states of the predicted perturbed cell, respectively.

Combining all these variables that should influence the final ranking of the gene perturbation, the objective function *O* shall be as follows:

Where a larger value indicates a gene perturbation close to the desired Q with effective properties and a low value indicates low capacity to perturb a gene with respect to the other genes in the set.

**Proposal for Q (Optional Task)**

Ideally, the cell state proportion vector should contain more active T cells than weak T cells. *Q* should be chosen for cancer immunotherapy such that tumor-killing T cell activity and growth rate is maximized, since we want active T cells to be able to proliferate. Activity is reflected by the proportion of T cells in progenitor and effector states. Meanwhile, growth and cell renewal, where a T cell performs cell division to proliferate, is represented by the proportion of T cells in cycling state.

However, a significant imbalance between activity and growth is undesirable in maintaining efficient tumor-killing activities. For instance, a sum of the progenitor and effector cell state proportion of 0.95 and a 0.05 of cycling state proportion means that a T cell has a low tendency to proliferate despite being an active tumor-killer. Conversely, a T cell with a low proportion of activity e.g., 0.1, but a high proportion of cycling state e.g., 0.9, infers that a cell undergoes rapid cell division of weak T cells which are those in the terminal state. However, we ideally prefer a larger ratio of progenitor to effector cell states. Overall, it is desirable that a good balance of activity and proliferation is demonstrated by a T cell in order to be considered effective in eradicating tumor cells.

Considering that we want the T cell to have a high probability of being in an active state, the sum of progenitor and effector state proportions should be large enough that this gives considerable confidence that a cell is an active tumor-killer.

Given that *Q* = [a, b, c, d, e], *one example* of an ideal desired *Q* may consist of the following proportions:

[0.6, 0.1, 0, 0.3, 0]

Let Active *A* be the sum of progenitor and effector state proportions, and *O* be the proportion of inactive cells i.e., terminal cell state proportion. Let *G* represent growth rate and cell renewal as given by the cycling cell state proportion. As a constraint, the cycling state should at least comprise 0.05 of the proportions. Therefore, we formulate a mathematical equation that represents the goodness of a *Q* for cancer immunotherapy applications.

where *max*(*a, b*) represents the maximum value between *a* and *b* to take into account a preference for the progenitor state *a*, while *max(G, 0.05)* puts a constraint to the minimum value of *G.* This equation gives a score between 0 and 1, with 1 indicating that the sum of *a* and *b* is equal to 2 times the maximum value of *a* or *b* and equal to or greater than *G*. On the other hand, a score of 0 indicates that G is equal to or greater than the sum of *a* and b. The equation penalizes huge imbalance between G and A by dividing by a value that approaches 0 as G becomes much greater than *A*.