Max score 6

676583:

The documentation is well written and clear. However, for the first formula, it is totally based on variance. From my perspective it is not an ideal indicator. Some disadvantages include: 1) take all the genes’ weights equally 2) simply take states into account by subsetting and calculating variance sums in each subset

The 3rd part scoring function for cell states looks nice and practical, with a “favor strength”. Review score: 5

675876:

The method is not practical and lacks mathematics or statistical reasoning. There is some denotation issue in the formula provided (pasted below).

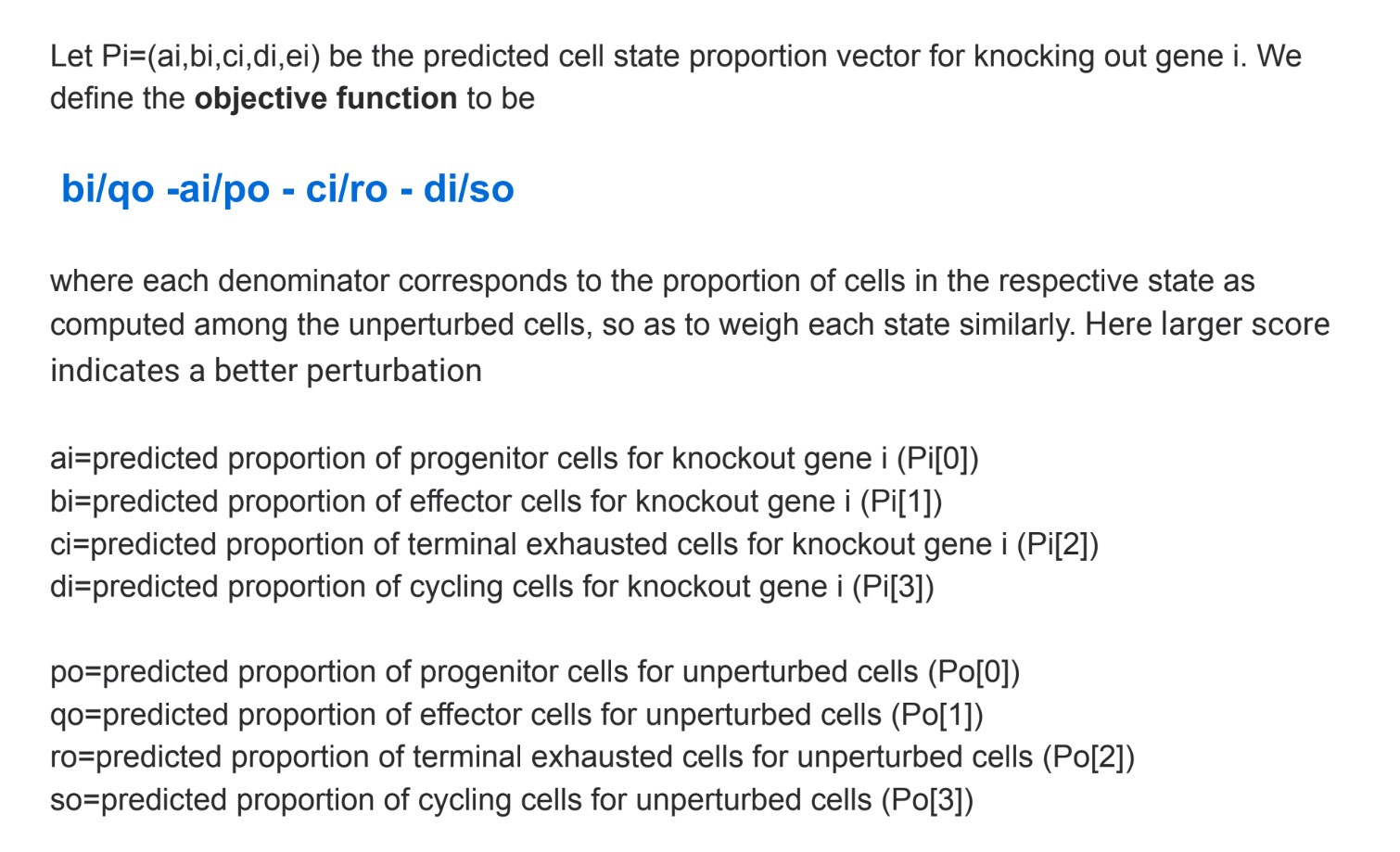
“Let "Po" denotes the empirical gene expression distribution of the unperturbed cells, i.e., "Po" is a distribution in 15,077-dimensional space. Similarly, let "Pi" denote the gene expression distribution of the cells obtained by knocking out gene i. Let Q denote the desired cell state proportion vector, i.e., Q is a 5-dimensional vector of probabilities that add up to 1. As an optional task, you are invited to submit your proposal for how to choose Q for cancer immunotherapy.”

“Po = Po-Pi”

Review score: 3

676613

The method simply uses effector proportion as an indicator and treats others as “harmful”. The method is clear and simple. It can be a good baseline.



Review score: 3

676646

Well written and feasible. The statistics part equation and procedure is clear.

Scoring function and formulation seems feasible. It can be a potentially good baseline, and have subtle connection to certain widely used graphical models, e.g. LDA. Review score: 6

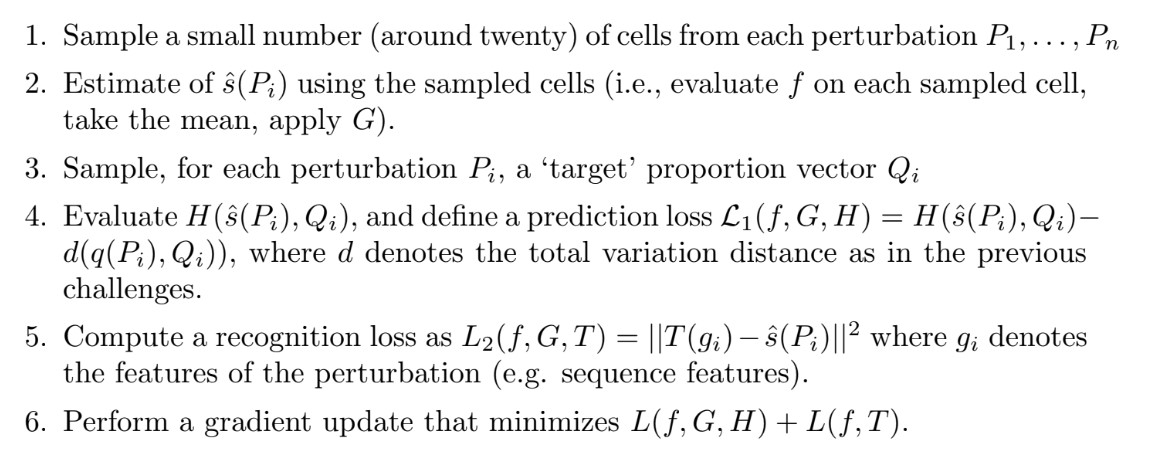
676651

Introduce Mahalanobis distance and cosine similarity to measure distribution differences. The formulation is clear.

Review score: 5.5

676628

Propose a method to sample gene expression data and compute loss to optimize the neural network. Technical details and results are provided. One disadvantage can be lacking explanation on how to handle sample size variances for cell states, perturbation samples and unperturbed samples.



Review score: 6

676333

The submission is not in the required form. It proposes a KL based method to measure the power of perturbation. The formulation in the text file is not clear enough to be judged as feasible, though using KLD makes sense according to the explanation.

Review score: 4

# Submission 674631

**Overview**: The authors propose (a pseudo-inverse of) a weighted average of knockout cell state L1-loss minus the unperturbed cell state proportions L1-loss. The weights are defined by the variance of cell states.

|  |  |  |  |
| --- | --- | --- | --- |
| Clearly explained: yes | Well-motivated: kind of | Novel features: yes | Computationally feasible: yes |

**Strengths**: The authors try to take into the variance of the data. The measure takes into account the unperturbed cell state proportion.

**Weaknesses**: Using variance without considering the sample size is a bad idea – small samples can have very small variance by chance. The measure is not intuitive, has no statistical interpretation, and the authors do not provide the metric’s range of values. It is also hard to judge how such a measure would work in practice, and the authors do not provide any examples.

# Submission 676621

**Overview**: The authors propose to create a KNN graph of cells to analyze neighborhoods and perform neighborhood enrichment analyses. Moreover, instead of comparing (enriched) cell state proportions directly, the idea is to compare delta vectors from the unperturbed proportion.

|  |  |  |  |
| --- | --- | --- | --- |
| Clearly explained: no | Well-motivated: no | Novel features: yes | Computationally feasible: yes |

**Strengths**: The idea tries to provide a directional component to the metric and compares the unperturbed-desired vector against the unperturbed-knockout vector.

**Weaknesses**: The description is vague and without references, making it easy to misinterpret the idea. Clustering cell data will make the sample sizes even smaller than in the original problem definition, which can create noisy vectors. Moreover, the method does not incorporate any measure of statistical significance or uncertainty. Finally, the method, when repeated for all potential knockouts, may be potentially computationally demanding, albeit feasible.

# Submission 676613

**Overview**: This paper focuses on the biological aspect of the problem and proposes ideal cell state proportions for cytokine and monoclonal antibody therapies.

|  |  |  |  |
| --- | --- | --- | --- |
| Clearly explained: yes | Well-motivated: yes | Novel features: no | Computationally feasible: yes |

**Strengths**: The writeup provides biological motivation and references. Focuses on the bonus task of defining ideal proportions for therapies.

**Weaknesses**: The proposed metrics do not take into account sample sizes or uncertainty. The metrics are basically a variation of the metric used for Part B of Challenge 2. The ideal cell state proportion for the cytokine therapy does not sum to one (that might have been intended, but I believe this can be a little troublesome in practice).

# Submission 676658

**Overview**: The authors propose summarizing gene expression distributions *P* with only the top *d* differentially expressed genes. This way, the authors base their scoring function on dx5 matrices rather than entire expression matrices *P* or only summary vectors *Q*.

|  |  |  |  |
| --- | --- | --- | --- |
| Clearly explained: yes | Well-motivated: yes | Novel features: yes | Computationally feasible: yes |

**Strengths**: The authors try to take into account and summarize *P*. By using a T-test for differential expression, the authors have a statistical tool for significance and a directional metric showing whether the expression of a given gene is in the right direction. The authors use the unperturbed cells as a reference. The authors also provide code and show how their score relates to L1-loss.

**Weaknesses**: The measure is not intuitive – it is hard to interpret the score values. Although the proposed score correlates with L1-loss, it can still behave weirdly – for one of the examples, the gene with the highest possible L1-loss got the highest score. Using the top *d* differentially expressed genes may be prone to correlation issues, i.e., the top differentially expressed genes may consist of a cluster of highly correlated genes from the same region (hotspot).

# Submission 676643

**Overview**: The author proposes to use Gene Set Enrichment Analysis (GSEA) results to predict cell states and use a softmax to turn that into cell state proportions. The predicted cell state proportions are then compared with the desired proportions via cosine loss.

|  |  |  |  |
| --- | --- | --- | --- |
| Clearly explained: yes | Well-motivated: yes | Novel features: yes | Computationally feasible: yes |

**Strengths**: The author tries to take into account the pathways related to the knockout gene. Instead of a user-defined set of cell state related genes, GSEA uses a p-value. The author provides code and intermediate results.

**Weaknesses**: The measure does not consider the sample size or uncertainty. In my opinion, using cosine loss alone can focus on the correctness of the direction without taking into account the distance to the desired cell state proportion. If the direction is perfect, but the perturbation effect (vector length) is very small, the knockout will be scored perfectly even though the effect is very close to the unperturbed proportion.

# Submission 676614

**Overview**: The authors propose variations of metrics defined for Part A and Part B of Challenge 2. The difference lies in the scaling of the components of the metrics–instead of normalizing by the unperturbed cell state fraction, the authors propose to normalize by the cell state fraction of zero-expression cells of the knockout gene.

|  |  |  |  |
| --- | --- | --- | --- |
| Clearly explained: yes | Well-motivated: yes | Novel features: kind of | Computationally feasible: yes |

**Strengths**: Simple metrics, which are close to those defined in Part A and Part B. The measures use the *P0*.The authors provide preliminary results and code.

**Weaknesses**: The values of the measures are not easily interpretable and do not take into account the sample size, statistical significance, or uncertainty. The method heavily relies on *P0,* and, in turn, some genes can have it easier or harder to perform according to this metric.

# Submission 676654

**Overview**: The authors propose a statistic that resembles the absolute log odds ratio to calculate distances between cell state proportions and scale that distance distances by the sample variance.

|  |  |  |  |
| --- | --- | --- | --- |
| Clearly explained: no | Well-motivated: no | Novel features: yes | Computationally feasible: yes |

**Strengths**: The authors try to take into account variation in the samples. The authors propose a non-linear distance function to compare knockout and desired cell state proportions. The “bonus” idea about performing several experiments over time is valid, albeit not that new and very costly.

**Weaknesses**: The description is very vague. It is hard to understand how the measure would behave in practice and why the authors chose this formulation. Moreover, the measure does not take into account sample sizes; hence, the variations alone may provide noisy results. Also, the measures do not use the unperturbed cell state proportion as a reference.

### **676666**

The scoring function is simple but logical. Correct use of mathematical notations and arithmetic with respect to the variables is observed in the formulated equation. A variety of variables that influence a perturbation’s final scoring was taken into account such as cell growth and difference of the perturbed state vector from unperturbed state and desired Q. However, explanation of the equation is lacking. It was not explained thoroughly the rationale behind getting the loss *L1* of *s(Po)* and *s(Pi)* and *L1* of Q and *s(Pi)*, as well as the method to obtain the cell count. While *s(Pi)* was assumed to be the predicted cell state vector from Challenge 1 and 2, the statistic does not take into account uncertainty and classification boundaries.

### **676643**

The author presented an interesting approach of summarizing Pi in terms of the significant pathways influenced by a gene perturbation. Moreover, as a distance metric to compare the predicted *Q* and desired *Q*, cosine loss was used, and the author was able to reference papers and argue why this is the best metric when it comes to analyzing classification accuracy. It is also greatly appreciated that the code written by the author is presented in the Appendix for ease of understanding of the readers. However, the scoring function does not include other variables that represent the effectiveness of a gene perturbation, which may include change from the gene expression of the baseline / unperturbed cells, variation in sample sizes, and growth rate. It was also not sufficiently explained how *Q\_hat* could be obtained from the formula presented by the author *:= softmax(−s(Pi)/10).*

### **676642**

The authors presented the statistic in terms of the change in cell state composition from the unperturbed cell state vector and added a binary variable *x\_i* to the statistic representing the significant change induced by the knockout. This statistic was then used as input to the scoring function together with the Q\_o obtained from Po and the desired Q. The addition of the binary variable x\_i is an interesting approach, where the authors propose defining a threshold based on the top 20 unperturbed cells that show a huge ΔQ. Aside from this idea, we also commend the technique they proposed where normalization is first applied before calculating the Euclidean distance. Despite these novel approaches, the authors were not able to take into account cell growth as well as classification boundaries While it is important to take into account the change in cell state proportion vector from the baseline state after a knockout, other variables should ideally be included to get a more holistic ranking of the gene perturbations.

**676635**

The authors used a simple and computationally efficient metric based on Euclidean Distance. Their approach recognises the drawbacks of L1 loss in this application and provides score normalization. However, the approach only considers the desired proportion vector Q and does not consider the summary statistic *s(.)* and by extension the gene expression proportions *Po* and *Pi*. Unfortunately, the scoring function was only based on the distance from the desired Q, and other variables were not included such as cell growth, classification boundaries, and uncertainty from number of cell samples

**676654**

The authors presented a detailed approach to calculating the summary static by considering the change in variance of gene expression through perturbations. This is useful as perturbations that do not change the gene expressions are unlikely to change the target proportions significantly. This also provided a method to differentiating distributions that are functionally different but scored similarly — a drawback of simply using L1-loss. With respect to Q, the approach recognises that an ideal distribution would be determined experimentally and may change over time. Perhaps due to this, an ideal Q was not explicitly proposed.

**676619**

The authors presented a detailed approach combining MAST enrichGO and L1 loss to calculate the perturbation score of a gene knockout. Their approach incorporated softwares/libraries such as CellOracle and scPred to produce computationally achievable cell state proportions. Moreover, it was justified why using L1 loss as the distance metric is better than other metrics. Predicted growth rate of cells from perturbation was also incorporated by using MAST. While their approach offers a computationally efficient method of comparing cell state vectors between perturbed and unperturbed states by utilizing existing libraries, uncertainty in classification boundaries were explicitly not included.

**676673**

The author proposed a variety of variables to be included in the scoring function and mentioned methods to achieve them. However, the author didn’t present the proposed equation nor clearly showed the idea in a way that readers without expert mathematical background could understand. Unfortunately, it was difficult to digest the proposed idea without the author showing the step by step process of how to obtain the final scoring function.

[Submission ID: 676661]

This competitor suggested 3 possible metrics for the scoring function: Kullback–Leibler divergence(KLD), graphical neural network, and dynamic system based metrics. A concrete scoring function have been proposed for Kullback–Leibler divergence based metric case, which reflects the difference of distribution of perturbed(Pi) and unperturbed(Po) gene expression. The competitor have demonstrated a clear contrast of the distribution of perturbed gene expression (Hif1a,Nr3c1,and Arid5b) to that of unperturbed cases. Since this scoring function has been based on the competitor's submitted models in other challenges, it may be relatively easily implementable. On the other hand, more detailed explanations such as how to obtain the distribution plots from zero-inflated data and summarized it as a statistics, figure annotations, axis labels, etc might be helpful for the reviewer to understand the proposed score function better. Also, as the competitor admitted, the proposed scoring function has not taken into account several aspects of desiderata requested by challenge3, such as the different sample sizes for the perturbations, different growth rates, different cell numbers, the classification boundaries of each cell state, etc. It would be also helpful to present any concrete form of scoring function for the other cases (graphical NN and dynamic system based metrics) for helping the the reviewer understand the competitor’s ideas better.

[Submission ID: 676657]

This competitor team has proposed a distance based scoring function of gene perturbations with the considerations of other factors such as growth rate(reflecting the factors of number of cells & cycling state) and uncertainty in classification boundary. They have nicely observed a correlation between number of cells and cycling state by calculating spearman correlation from the dataset, then relate it with the growth rate, and incorporated it into the scoring function. The absolute difference between desirable and undesirable states has been devised as a penalization factor to reflect the uncertainty of classification boundary. Considering the components of the proposed scoring function, its implementation should be no issue. On the other hand, some edge cases need to be addressed to improve the proposed scoring function. For instance, when calculating the distance between the perturbed cell state proportion Q\_hat and the unperturbed Q\_bar(∆Qio = Q\_hat − Q\_bar), a large value of ( c i −c o )^2 + ( e i −e o )^2 can cause misinterpretation of the resulting large value of the ∆Qio as a preferable state, which is actually not the case. Converse edge case is also true for the distance between the perturbed cell state proportion Q\_hat and the desired Q(∆Qi = Q\_hat – Q\_desire) metric. Also, in the penalization formulation of s(Pi)=1− |a+ b – c|, when c >> a+b, the less penalization by s(Pi) is not appropriate while it should be more penalized in the scoring function.

[Submission ID: 676612]

This competitor has presented an algorithm of assigning cell states of the perturbed genes by comparing similarity and dissimilarity of the rest of 15006 genes in both perturbed and unperturbed conditions, which seems to be implemented for the competitor’s submission of the other challenges. The provided diagram is helpful to understand the context/setup of the procedure. Stepping through the procedure has elucidated how a gene expression profile in both conditions has been utilized by calculating a correlation with that of the perturbed genes so that both Pi and Po could be implicitly utilized in the process instead of presenting an explicit statistical summary of Pi and Po. On the other hand, how to obtain Ranks/probability-score for cell state vector of perturbed genes need to be elaborated with the considerations of other desiderata of the this challenge such as:he different sample sizes for the perturbations, different growth rates, different cell numbers, the classification boundaries of each cell state, etc. In addition, in order that this algorithm should be effective, the enough knockout/down of gene expression level by PerturbSeq experiments should be warranted, but there seems to exist a quite a wide variability in the expression level in perturbed gene data. In some cases, the expression level is even higher. For instance, 'Tcf7' shows more gene expression level in the perturbed condition than unperturbed condition in dataset. This fact could weaken the hypothesis in step-d:"The intersection of most dissimilar genes in step-c and similar genes in step-d can be considered as genes of similar type". Then, checking the different gene expression level of the perturbed gene (such as knockout/down vs. stay-the-same vs. up-regulation) and handling it accordingly would be a way to mitigate the issue.

[Submission ID: 676646]

This competitor team has impressively well addressed most parts of challenge requirements. Data analysis in section 1 is insightful and informative. Based on this analysis, both a statistics summarizing Pi and a scoring function with the inputs of s(Pi), Po, and Q have been proposed with the thoughtful considerations of the other aspects such as different sample sizes for the perturbations, different growth rates, different cell numbers, the classification boundaries of each cell state. Code and plots included in the submission shows that necessary components of the statics and the scoring function are already implemented. Overall, it seems that his competitor team have done a nice work and presented a strong proposal. By the way, one thing I noticed in the statistic proposal part is that they have excluded cells that have higher target gene expression than the mean of the unperturbed cells(related to the observation in section 1.5). Is this really a necessary step? My naive thought (as an ignorant of PerturbSeq experiment) was that the given dataset had passed the quality control by the challenge host, and the gene expression level variations in both perturbed and unperturbed cells of dataset in basal condition might not really reflect the gene expression capability of the gene in a stimulus condition, which we may not know until we put both the perturbed and the unperturbed cells in the same stimulus condition and directly compare respective gene expression levels responding to the stimulus. if I am wrong about this, please safely ignore this opinion of mine.

[Submission ID: 676658]

This competitor team has presented summary statics of Pi and Po which is based on the expression values of deferentially expressed top-n genes for a cell state (one-vs-rest comparison), and used the weighted sum of log ratio of the Pi and Po statistics as a main component of the proposed scoring function. They have nicely demonstrated that the proposed scoring function is correlated with L1-loss, and provided the some useful intuitions of the proposed scoring function as well as the implementation python code. Also, they have addressed the issue of the uncertainty of a predicted perturbation by considering confidence interval and suggested some possible implementation methods such as training ensemble of randomized predictive models. On the other hand, other factors such as different sample sizes for the perturbations, different growth rates, and different cell numbers need to be incorporated into the current proposed scoring function.

[Submission ID: 676628]

This competitor has presented a complex but powerful proposal. It consists of 4 neural networks:f(gene expression program predictor), G(multi-cellular program predictor), H(gene perturbation capacity predictor), T(gene expression program predictor by gene sequence features). Presented code and plots have nicely demonstrates that these neural networks can learn statistical features of dataset. As a result, f and H can be confidently used a summary statistics of Pi and a scoring function respectively, and furthermore the incorporated recognition network H looks like an interesting component in the proposal, which uses an external language model embeddings of the gene sequences to predict gene expression (or multi-cellular) program by the gene sequence features only. The different sample sizes for the perturbations also have been considered by restricting the number of samples in training phase for each perturbation. Although a few parameters such as different grow rate and class boundary still need to be considered to be incorporated, it is a minor point. Overall, this proposal looks interesting and promising.

[Submission ID: 676672]

This competitor has presented a distance based scoring function. The provided notation table and presenting format clearly delivers the context/setup of the proposal. The summary statistics of Pi has been chosen as S(Pi) = Q-hat, which may be the prediction model implemented by the competitor in the other challenges. Then, a bounded ratio of distance metric has been presented as a scoring function, which can be generalized with any distance measure such as Euclidean distance & KL-divergence. On the other hand, other aspects of requirements such as the different sample sizes for the perturbations, different growth rates, different cell numbers, the classification boundaries of each cell state, etc need to be incorporated. Also, for the distance based metric, the distance of c\_i & e\_i states(undesired states) in the metric needs to be handled differently from the distance of a\_i,b\_i,&d\_i states(desired states) because large distance of c\_i&e\_i can lead to the misinterpretation of the corresponding large score.

**Review 1:**

Submission ID: 674189

The proposed solution entitled: Clonally Aware Measure of Effector Longevity or CAMEL is an interesting approach to the challenge. Optimizing the overall behavior of clones rather than cells makes sense. The author proposes to apply this by measuring the clonal capacity not as a function of time, but as a function of clone size. The theory is worth testing in my opinion. However, the proposed linear regression between clonal expansion time and terminally exhausted fraction is more of a baseline or a start. Further research should be explored to find the best way to solve or compare both variables.

Although the author has a novel approach for the challenge, he/she missed providing a clear statistic and scoring function as per the requirement of the challenge.

**Review 2:**

Submission ID: 676333

While going through the submitted files, I found it to be quite confusing as to where we should be reading and for what objective. The submission had 4 files: Statistic.txt, research.txt, Research\_2.txt, and the goal of the scoring function.txt. As for the statistic part, the author proposes to use the Euclidean distance to measure the error for the gene expression distribution between the knock-out and the desired cells. The distance would be used to train a couple of ML algorithms along with cross validation to ensure generalization of the model to new data. While this approach is valid, I don’t see novelty in the method.

However, calculating a weighted average with penalty to measure the difference in the distributions for the scoring function is a valid idea. The formula provided for the scoring function using the Kullback-Leibler divergence is a good proposition.

Note: There were no references / citations given in this proposal.

**Review 3:**

Submission ID: 676583

The statistic proposed by filtering the expression through their mean, variance across cell states, and variance across knocked-out cells is valid. However, the approach is not unique and needs experimentation to see if it will yield into the desired results.

Th proposed scoring function for perturbations is based on a numerator that takes the mean expression of the genes across the cell states, divided by a denominator that takes the variance of the gene’s expression across the cell states. The function integrates a variance variable Vj that decides the cell state boundaries’ confidence, which I find interesting and novel as an approach. In addition, it integrates a co-variance variable for all the states to favor the desired cell sate. Without testing the formula, it is difficult to judge its effectiveness, but it looks promising.

On the other hand, the author proposes a scoring function for cell states. It is not clear how the formula came about or what was it based on. This formula needs further explanation.

As for the proposal for Q, the author argues that the T cell states are currently independent of each other, only to refute this this idea later. However, it is clear from the beginning that the states are not independent from each other. Beyond this idea, the proposal for Q is not clear.

**Review 4:**

Submission ID: 676613

The author presented us with a detailed and well summarized steps of cancer definition, certain treatments, and cell behavior. However, I believe that the author is deviating from the challenge request. By proposing a different treatment, Cytokine Therapy, the author is venturing into an unknown field especially for us as data scientists. This is an approach that was not required in the challenge. We don’t know whether favoring the effector and cycling cells instead of the progenitor cells would reach to a better result. Moreover, the author didn’t propose a statistic nor a scoring function that were clearly defined as per the challenge request. This is not to state that Cytokine Therapy is ineffective, but rather to say that it is not the requirement of the challenge.

**Review 5:**

Submission ID: 676622

I couldn’t find any document that has a write-up that solves challenge 3 requirement. The files provided were .py files for challenge 1 & 2. The Text documents were a recap of the 3 challenges requirements only.

**Review 6:**

Submission ID: 676646

The proposed statistic is valid and interesting. Using Principal Component Analysis to capture the variance in the gene expression, then to predict the cell type label, and then to predict cell type probability vector for each perturbed cell is a smart approach. After excluding cells with higher target gene expression than the mean of the unperturbed cells and averaging the probabilities of the remaining cells, the authors concatenate the average cell type probabilities with the total number of cells per perturbation. This statistic has a high potential.

The authors propose a scoring function with a weighted sum of the logarithm of the significance of the change induced in cell type numbers multiplied by 1 or -1. While this function is valid and promising, the authors didn’t elaborate well on how to handle the weights, increasing or decreasing. Suggesting an increase in the weights based on previous observation is not enough. Moreover, the authors suggest normalizing the observations to a constant of 100 cells, so that the desired signal in the distribution is not lost. I find that flattening the observations to a constant number is limiting, especially when there isn’t a method to determine this number.

Note: There were no references / citations given in this proposal.

**Review 7:**

Submission ID: 676672

The author proposes a statistic like the one provided in the challenge. Therefore, no novelty was introduced to this part of the challenge.

As for the scoring function, the author proposes the Euclidean distance which is the ratio between the ideal state proportion vector Q and the predicted Q-hat. While this method is a valid approach, it doesn’t bring a unique or innovative approach to this challenge. Moreover, this approach doesn’t take into consideration neither the growth factor nor the uncertainty factor.

Note: There were no references / citations given in this proposal.

# 676661

The author proposes a scoring function defined as the Kullback–Leibler divergence (KLD) between the unperturbed and perturbed gene distributions. The function is predictable for unseen perturbations as it is defined in terms of dataset genes. The usage of the KLD is motivated by interesting observations about the gene distribution similarities under the KLD. However, the scoring function does not consider changes in the desired cell state proportion vector Q. Moreover, the scoring function does not consider changes in any predicted statistic nor does the author propose a statistic as required by the challenge-3 specifications.

# 676612

The author provides an interpretable and clearly explained estimate for cell state distribution as the proposed statistic. However, the author does not propose an accompanying scoring function as required by the challenge-3 specifications. Nonetheless, the proposed scoring function uses uniquely uses the cells as features and genes as the samples, which goes beyond what is used in challenge 2. The method appears biased toward predicting the highly represented cell states in the dataset. Namely, the statistic classifies a perturbation gene “by considering the maximum value of all cell states” for a select group of dataset genes, which inherently favors the highly represented cell states.

# 676657

The authors propose a scoring function composed of *sub-metrics* incorporating several intuitive observations they make about cell growth rates and distances between cell state distributions. The scoring function favors perturbations with large growth rates in perturbed cells providing evidence of its strong correlation with desirable cell states. Namely, they uniquely incorporate growth rate as a *sub-metric* in the scoring function. However, further explanation is needed regarding how the scoring function scales the *sub-metrics* with respect to each other.

For example, it is not immediately obvious nor explained how the growth rate should be compared/scaled against the proposed statistic since it is a regulatory term unrelated to the growth rate. Moreover, a similar relative scaling problem seems to arise in the proposed statistic, where the effector and progenitor states are presumed to be equally beneficial/weighted without justification.

## 676672

The author proposes a clearly explained scoring function whose output is interpretable. The proposed scoring function is computationally ideal for large datasets given its simplicity. However, further investigation (above a single example) needs to be presented to satisfactorily motivate the author's choice for Euclidean distance over Kullback–Leibler divergence or other reasonable alternatives with very different properties (for example the MAE, Cosine distance, etc.). The author does not offer a specific and clearly defined proposal for a statistic that summarizes the gene expression distribution as required by the challenge-3 specifications.

# 676642

The authors put a relatively high amount of effort into clarifying the details of their proposed statistic and scoring function, including finding external research that motivated/informed parts of the scoring. The uniquely defined statistic is centered around defining a binary regulatory feature x that can potentially integrate biologically relevant information. However, I have concerns that the scoring function given the statistic will be uninformative (x=0) for the majority or substantial amount of possible perturbations. Namely, the threshold for when x=0 is determined by “the top 20 percentile of ΔQ values” computed over unperturbed cells or the entire dataset. The Euclidean distance is the basis for the presented scoring function, which equally weighs the differences between the individual elements, giving potentially undesirable scores. For example, if a perturbation is estimated to be terminally

exhausted (i.e. 𝑄𝐴 = (0, 0, 1, 0, 0)) will identically be ranked by the scoring function to a

perturbation that is estimated to be in a more desirable effector state (i.e.the given desired proportion vector is the progenitor state (i.e. 𝑄𝐵 = (0)., 1, 0, 0, 0)) if

𝑄 = (1, 0, 0, 0, 0)

# 676621

The proposed scoring function and statistic both rely on performing enrichment analysis, a more principled approach that better interprets the differences between clusters of cells. In particular, their statistic requires using MILO (see “Milo: differential abundance testing on single-cell data using k-NN graphs” Dann et al. (2020)) to obtain their statistic, transformable into an estimator of the cell state proportions. MILO is a computationally efficient framework for performing large-scale enrichment analysis. Thus the proposals appear computationally feasible. The authors introduce a new quality control step by excluding underrepresented perturbations certainly improves the overall quality of the data and ultimately improves the statistic and scoring functions derived from it. However, the proposed statistic is potentially biased toward highly represented cell states found in the dataset. Namely, the statistic relies on categorized neighborhoods (of cells) whose categorization is based on a cell state “majority vote” or the state of an “index cell”, both of which will likely favor the majority state. Additionally, the scoring function is defined in terms of discrete increments of +1 and -1, which will likely result in many ties produced by the scoring function producing for a fixed desired cell state proportion vector Q.

## 676622

The author provides a concrete understanding of essential aspects of the challenging task. However, the author does not propose a statistic as required by the challenge-3 specifications. Nonetheless, the author provides a scoring function proposal that is very concise and simple and is therefore computationally feasible for large datasets. However, the author needs to give much more detail motivating and clarifying the specifics of the proposed scoring function. For example, the author needs to provide more details justifying using the AUC (Area under the ROC Curve) in the scoring function which is not defined for non-integer label values, the desired cell state proportion vector Q.

**Review for submission 676635:**

The strength of this proposed scoring function is that it is simple and should be easy to implement. The main “novel” feature is the replacement of the L1 loss function with Euclidean distance. The submitter gives some, but probably not enough motivation for why Euclidean distance is better than L1 loss. A major weakness is that none of the desiderata are accounted for in the scoring function. Finally, for the optional task, the proposal given is not of a desired value for Q.

**Review for submission 676628:**

This proposal seems to have a lot of potential, although I am not very familiar with the methods used. The statistic is clearly defined, but the scoring function is not. While not explicitly stated, the scoring function should involve the sum of the prediction and recognition loss terms which are defined. It is not clear which of the desiderata are accounted for. Although I am not certain of how good a fit this approach would be, it warrants a further look by experts.

**Review for submission 676643:**

This proposal is well written and has many strengths. Both a novel statistic and scoring function are provided. The statistic using GSEA and scoring function using cosine loss are clearly explained along with the motivation for their choice. Evidence is given for how this is superior to the scoring techniques used in Challenge 2. The proposal includes some of the desiderata, but not all. A straightforward code implementation is also provided.

**Review for submission 676634:**

Only an ipynb Jupyter Notebook file is provided. There should be a writeup in markdown, pdf or word format. Looking at the ipynb file, the content is about predicting cell state and ultimately measuring the accuracy of a perturbation prediction which is not the goal of Challenge 3. No clear statistic or scoring function is given and therefore it does not fulfill the requirements.

**Review for submission 676672:**

This is similar to 676635 with a couple minor added strengths. A strength of this proposed scoring function is that it is simple and should be easy to implement. The main “novel” feature is the replacement of the L1 loss function with Euclidean distance. An additional strength is the inclusion of P0 dependence in the formula. However, none of the other desiderata are included. A proposal for Q is given with some justification, but probably not enough. In general, not enough motivation for this approach and its advantages over existing methods is given.

**Review for submission 674189:**

This submission has some interesting ideas about how knocked-out genes may result in a cell state proportion with high terminally exhausted fraction and still be effective. This may be because the process of knocking-out the gene induces a larger clonal expansion rate. Although this might’ve been a nice paper for expert review, there is unfortunately no concrete statistic or scoring function.

**Review for submission 675654:**

This document basically describes selecting the winner. This is after the selection of the top 20 participants based on the scoring functions. Unfortunately, the actual derivation of a statistic or scoring function for part C is not described. Therefore, this submission does not fulfill the requirements.

# Submission 676658

Pros:

* The algorithm is relatively simple and justified
* The score shows the direction of change from the ***Q0*** distribution

Cons:

* Authors do not consider classification boundaries
* Before computing the score, the parameter ***d*** of statistics ***s()*** has to be specified. The authors suggest the value of the parameter ***d* = 5**, however, it is not clear on what basis this value was chosen, and what value of the parameter ***d*** is optimal.

# Submission 676620

Pros:

* The score function is well described and justified
* The score function using the well-described formulas like Cohen’s Kappa statistics or Hoeffding bound adapted to the specific problem Cons:
* 𝞳***TL*** consider only lower confidence bound of 𝞳***T*** which can be treated as a worst case scenario score.
* The authors did not directly present how to calculate/predict the distribution of ***Qi*** (***s()*** statistics) however, they dedicated a chapter to a discussion of decision boundaries. Therefore, I assume that the authors relied on a previously trained classifier to determine the distribution of ***Qi***

# Submission 676670

Cons:

* The proposed calculation of cluster density is not normalized and depends on the number of samples.
* The description of the algorithm lacks labels and equations, which makes it difficult to reproduce the process.
* If I understand correctly, the score function is defined as the proportion of desired cells to all cells adjusted for the quality (density) of the set. Such a function does not take into account the uneven distribution of cell type Q.

# Submission 676619

Pros:

* The scoring function is simple and and well described
* In my opinion, using Gene Ontology terms in a scoring function is novel and interesting idea Cons:
* Do not consider uncertainty
* Implementing ***g(Po, Pi)*** in the manner presented in the report may lead to an underestimation of the entire score function as the gene ratio decreases. In particular, when the gene ratio is close to zero (the authors have secured by completely zeroing out the score function)

Submission 675654

Specification of the score function and s statistics is not included in this report

# Submission 676634

The report is actually a jupyter notebook with an ML model implementation for cell state classification. The scoring function is not included.

# Submission 674189

The specification of the score function and s statistics is not included in this report.

Instead, the author proposes a different way of scoring cells with the knocked out genes based on clonality (size of clones), not on cell type distributions. Because I do not have a biological background, I do not want to attempt to evaluate this proposal. **I therefore suggest that the paper be read in its entirety and evaluated by an expert panel (e.g.**

**from Eric and Wendy Schmidt Center at the Broad Institute of MIT)**

# 676583

Strengths:

1. The proposal for Q has been clearly described and its usage makes sense.
2. The usage of variance and coefficients is well done.

Weaknesses:

1. The missing score code makes it difficult to implement the formulas provided, which are already quite sophis cated.
2. The provided sta s c is unclear, which may hinder the understanding of the data and limit its usability.

# 676666

Strengths:

1. The usage of the ra o f\_i/f\_0 is a smart choice to represent the change with respect to unperturbed cells.
2. The usage of L1 loss and normalized score calcula on by comparing with Q are effec ve ways to analyze the data.

Weaknesses:

1. The descrip on is very unclear, making it difficult to understand the context and the goals of the analysis.
2. Q is not defined, which hinders the ability to interpret the results of the analysis.

# 676333

Strengths:

1. The usage of Kullback-Leibler (KL) divergence is a good choice as it measures the dissimilarity between two probability distribu ons and provides a measure of the informa on lost when approxima ng one distribu on with another. KL divergence is commonly used in machine learning and informa on theory applica ons.
2. The penalty term Mahalanobis distance is a smart choice as it takes into account the correla on structure of the data and adjusts the distances between samples based on the covariance matrix. This can be useful for iden fying outliers or anomalous data points.

Weaknesses:

1. The forma ng is unclear, making it difficult to understand the proposed formula and how the different components of the formula relate to each other.
2. Q is not defined, which makes it difficult to interpret the results of the analysis or to determine the relevance of the proposed formula.

# 676673

Strengths:

1. The usage of Kullback-Leibler (KL) divergence is useful for measuring the dissimilarity between two probability distribu ons and can be applied to a variety of fields, such as informa on theory, sta s cs, and machine learning.
2. An uncertainty measure is useful as it can provide insight into the reliability or accuracy of the data or model, which can be important for decision-making.

Weaknesses:

1. Q is not defined, making it difficult to understand the context or goals of the analysis.
2. A clear formula is not provided, which can make it difficult to understand the proposed methodology or to reproduce the results.

# 676642

Strengths:

1. The sta s c func on is very clear and well explained, which helps to understand the methodology and the results.
2. The scoring func on is well researched and realis c, providing a useful tool for evalua ng the data.

Weaknesses:

1. A final summary would have been helpful, especially considering the length of the document, as it can help to provide a concise overview of the main findings and conclusions.
2. A sample code would have been helpful to demonstrate how the proposed methodology can be implemented and to facilitate its replica on and adop on.

# 675654

Strengths:

1. The response is short and concise, which can be useful for providing a quick answer or summary.

Weaknesses:

1. The response is not very detailed and does not appear to be well researched, which can limit its accuracy and reliability.
2. The response does not provide a mathema cal approach or any formulas or references, which can make it difficult to understand the underlying methodology or to verify the validity of the informa on provided.

# 676619

Strengths:

1. The response is very well researched and documented, with a clear and detailed explana on, which can enhance the reliability and validity of the informa on provided.
2. The response refers to useful materials and references relevant packages, which can help readers to further explore the topic and apply the methodology in prac ce.

Weaknesses:

1. The response may be lengthy, which can make it more difficult to read or digest, and can also require more computa on to implement the proposed methodology.

674189

The author of this writeup appears to have a background in biology, and their approach focuses on “metrics that select perturbations that are known to drive tumor control and clearance,” as opposed to the Immune Checkpoint Blockade Therapy approach, described in Challenge 2. Specifically, their approach focuses on PD-1 KO, which they claim “can arguably described as the ‘gold standard’ of cancer immunotherapy (with approved use for over 20 types of cancers),” even though PD-1 KO cells are not cell-intrinsically less prone to terminal exhaustion. The author’s proposed statistic appears to be the terminally exhausted fraction as a function of clone size, where a clone is defined as cells in a tumor with a shared gRNA-UMI. They did not explicitly describe their scoring function (or I was not able to discern it from the writeup). Thus, the main weakness of the writeup is the lack of detail in terms of their computational approach, following PD-1 KO; they “recommend a very simple, robust means of regression between [‘clonal expansion time’ and ‘terminally exhausted fraction’],” but it is not immediately clear to me how this would be incorporated in a scoring function. Overall, I believe the approach is biologically sophisticated, but can use more computational finetuning.

674631

The writeup clearly defines the statistic s(∙), but does not appear to define a separate scoring function. The explanation of the statistic s(∙) is fairly straightforward, but I am not fully following the motivation behind why/how they are defining certain quantities. For example, they define the variance of *Li* as the sum of the variances of each cell state proportion, and take the sum of all losses *Li* weighted by their variance *vari*, but this statistic does not give a value for a specific gene *i* knocked out, but instead will give the same value for every gene *i*. Thus, I am unsure of what this statistic actually tells one about any specific perturbation. In conclusion, the writeup is very straightforward and clear in its mathematical definitions, but in my opinion, does not provide enough biological or computational insight into why the statistic is defined as it is.

676612

The writeup succinctly outlines each step of their procedure in an easy-to-follow manner, and provides a figure at the top that defines all the gene expression matrices they use and their sizes. However, the motivation behind these steps is not explained; for example, I am confused about why they are choosing the most dissimilar genes between the knocked-out gene *i* and all non-knocked-out genes in Step-c. Also, in Step-e, they state, “Assign a cell state for *Genei* by considering the maximum value of all cell states,” which I do not understand. The procedure seems to assign a cell state for each knocked-out gene, but there is no explanation about how this ties into a statistic or scoring function (both of which there is no mention in the writeup). The approach is definitely computationally feasible, but does not seem to address the question that Challenge 3 is asking.

676613

The writeup begins with a background of cancer and cancer immunology, then proceeds to describing different existing therapies that maximize specific objective functions: cytokine therapy and monoclonal antibodies therapy, each with their own desired cell state proportions vector Q. The author clearly has a background in biology, as the writeup focuses on these existing therapies. A statistic and scoring function are not explicitly mentioned in the writeup, but two objective functions, corresponding to the two therapies, are provided, similarly to the CAR T-Cell Therapy objective function in Challenge 2. Thus, it appears to me that the author intended on tackling the optional task of how to choose Q for cancer immunotherapy in this writeup, but the actual task of defining a new metric for ranking perturbations for Challenge 3 is not addressed.

676621

The writeup outlines each step of their procedure with a decent amount of detail into how to finetune the procedure (e.g. setting a large enough K in KNNs such that there are < 20% neighborhoods that are predominantly unperturbed cells, and filtering out KOs that have fewer than 50 cells). They clearly define a statistic, but do not provide a separate scoring function. Also, I do not think the proposed statistic incorporates novel features beyond what was used in challenge 2, since they run KNN on both unperturbed and perturbed cells, thereby merging *P0* and *Pi*, instead of treating them separately. Overall, their approach is computationally feasible and seems fairly easy to reproduce, but they do not address the concerns regarding different growth rates and number of cells depending on the perturbation, and there is minimal discussion of the motivation behind their procedure.

676669

The writeup repeatedly refers to the previous Challenge submissions, of which I do not have access. They claim, “An important factor to use in the classification should be the number of close gene ontology affiliations that the given gene has in the GO database for *mus musculus*,” but classification is not a relevant task in Challenge 3. The writeup appears to be heavily adapted from previous writeups for Challenges 1 and 2, with minimal editing for Challenge 3. Furthermore, neither a statistic nor a scoring function is mentioned in the writeup, so I am not sure how to evaluate it, other than stating that using Gene Ontology correlations in Challenge 3 seems like a valid approach, but it is not described in any detail in this writeup.

676670

The writeup is well-organized into a “Problem Statement” section and a “Proposed Method” section with subsections for the statistic and scoring function. Their statistic is a “cluster quality metric represents whether our clusters [following data pre-processing and Leiden clustering] are both internally dense and maximally separated from each other,” which is a creative idea compared to just using the L1 loss from Challenges 1 and 2. Their scoring function takes this cluster quality metric and divides the estimated state proportion by it. I was a bit confused with this part, because the writeup does not clearly define what the scoring function actually is: they state, “If, for example, your objective function was to maximize the proportion of progenitor, effector, and cycling cells, and the estimated state proportion vector from your perturbation, *Qi*, was [0.2, 0.1, 0.3, 0, 0.4], the proportion of cells in your desired state would be 0.7,” which seems to imply that they are using the L1 loss, but this is not explicitly stated. Overall, the approach is computationally feasible and incorporates other aspects of the data beyond that of Challenge 2, but more detail regarding the motivation behind their statistic and scoring function would be appreciated.

# 676644

## Strengths

1. The proposal includes both a statistic and a scoring function to evaluateperturbations, which is well-aligned with the requirements of the challenge.
2. The proposal considers both gene expression and state proportion at theindividual cell level, which provides a comprehensive measurement of the perturbations.

.

1. The proposal takes into account the sample size of various perturbations, whichhelps to reduce the impact of contingency and improve the robustness of the results.

## Weakness

1. The proposal does not clearly explain how the final statistic (i.e., ʻ(·)) will beused as a scoring function to evaluate perturbations.
2. The sub-statistic for the measurement of state distribution is not well-motivatedand may not accurately reflect the desired state distribution.
3. The computational feasibility of the proposed method is not evaluated, whichmay lead to difficulties in implementation

# 676620

## Strengths

1. Clear definition and motivation of the proposed metric, Kappa-TVD-LCB (κT L), which considers the target cell state proportions, reference cell state proportions, and the confidence of the estimation.
2. The use of Total Variation Distance (TVD) as a measure of distance betweencell state proportions is a well-justified and reasonable choice.
3. Incorporation of Kappa statistic as a relative performance measure provides aframework for comparing the effectiveness of different perturbations.

## Weakness

1. The proposed method may not take into account other important factors such asthe time of measurement, growth rate, and number of cells, which may affect the accuracy of the results.
2. The use of TVD and Kappa statistic may be limited in their ability to capturecomplex relationships between cell state proportions, making it important to consider other metrics as well.
3. The proposed visualization method for analyzing and selecting the bestperturbations is not well described.

# 676630

## Strengths

1. The scoring function is clear and well motivated. It gives a score for each genebased on how it affects the state proportions of the T-cell populations, compared to the unperturbed state
2. The scoring function takes into account uncertainty by penalizing gene sampleswith fewer observations.
3. The proposal allows for flexibility in the weighting of each cell state, and in thedesired cell state proportion vector (Q).

## Weakness

1. The scoring function is computationally intensive as it requires summing over allthe T-cell states for each gene.
2. The weighting of each cell state is not specified, which can impact the validity ofthe results.
3. The scoring function does not take into account the classification boundariesof each cell state. This information could be important for evaluating the effectiveness of the perturbations.

# 676635

## Strengths

1. The proposal is based on a simple and well-known metric, the Euclideandistance, which makes it easy to understand and implement.
2. The proposal considers the spread of the error across different dimensions,which is a more informative metric than considering the error in one

dimension only

1. The proposed estimation method is computationally feasible and can be easilyimplemented.

## Weakness

1. The proposal only uses the Euclidean distance, which might not be the bestmetric for capturing the difference between the desired cell state proportion vector and the cell state proportion vector after knockout gene.
2. The proposal does not explicitly take into account the computationalfeasibility of the scoring function.

# 676634

## Strengths

1) The submitter has provided a .ipynb code for analyzing gene expression data to predict T cell states (effector, progenitor, and cycling) by ranking the occurrence of each gene in each state, and the code can be run on the top genes for the best state proportions.

## Weakness

1. The code does not include a scoring function or statistic which is an critical component of thischallenge.
2. Unfortunately, the submitter has not provided a solution for the assigned challenge and, in addition,the submission format file for Challenge 3 appears to be missing..
3. It does not incorporate any novel features beyond what was used inchallenge 2.

# 676642

## Strengths

1. Provides a clear and motivated explanation of the statistic andscoring function that it is proposing to use.
2. The framework is adaptable to any desired cell composition and takes intoaccount the initial transcriptomic state of a biological system, making it transferable to other cell
3. Computationally feasible as it reduces the search space to

interesting candidates by predicting the norm of the changes in cell composition rather than the effect itself.

## Weakness

1. The scoring function may not take into account uncertainty due to thedifferent sample sizes for the perturbations in the training dataset.
2. The scoring function may not be computationally feasible if it requiresprocessing large amounts of data.
3. Submission is limited in its ability to account for the compositional structure ofcell state fractions.

# 676622

## Strengths

1) The submitter has provided a .ipynb codes for different therapy to categorize gRNAs based on their associated cell states and store them in different lists.

## Weakness

1)The code does not include a scoring function or statistic which is an critical component of this challenge.

1. It does not incorporate any novel features beyond what was used inchallenge 2.
2. Challenge 3 submission format file appears to be missing.

**Review 1:**

Submission ID: 674631

The author didn’t clearly differentiate between the statistic from the scoring function proposal. However, we can infer that the author utilized the same statistic as the one proposed in the challenge without bringing any novelty or change to it.

On the other hand, the author proposes an interesting scoring function based on the inverse of a weighted average of the knockout losses minus the loss of the unperturbed distribution. Introducing the variance in this case gives less weight to the variables with most uncertainties and account for the growth rate. Taking the exponential of the denominator and the numerator, however, is not well justified to avoid negative values in the scoring function. This formula can be further explored.

Note: There were no references / citations given in this proposal.

**Review 2:**

Submission ID: 676621

The authors propose an interesting and novel approach to the statistic part of this challenge. Computing KNN graphs across cells and then utilizing neighborhood enrichment analysis to identify the strength of the perturbations is a smart approach. Furthermore, the approach is detailed in implementation through the use of GLM fits using the MILO package. However, the authors propose to count the number of neighborhoods based on the spatial FDR with a threshold of 0.15. There is no further explanation as to how this number came about. A 5-dimensional vector N is obtained by counting the number of enriched neighborhoods in each cell state, whereby then the vector is normalized to account for data imbalance. Overall, the statistic function is well thought of.

On the other hand, the authors propose a scoring function that aligns the N vector with delta Q by taking the negative and positive enriched neighborhood count as -1 & +1 denoted as D. Using the L1 distance, the score is calculated with a normalized D vector and the delta Q. Although that this is a valid approach, but the sample size, the growth rate, or the uncertainty in the data were not considered.

**Review 3:**

Submission ID: 676651

The author proposes a statistic based on a desired vector Q with proportions defined as (0.6, 0.25, 0.05, 0.1, 0.00). However, we don’t know how the author reached these desired proportions, and there is no explanation to support this strategy.

As for the scoring function, the approach is rather interesting. On one hand, the Mahalanobis distance is a good alternative to the standard vector distances such as the Euclidean distance because it accounts for correlations between the variables using a covariance matrix. The covariance matrix accounts for the uncertainties in the measurements as well as the growth rate. On the other hand, cosine similarity measures the similarity between the vectors while considering the magnitude and the direction of the vectors. The dot product of these two terms seems promising.

Note: There were no references / citations given in this proposal.

**Review 4:**

Submission ID: 676654

The author proposes a statistic that incorporates information about the distributional variety of gene expressions. The objective is to capture the differential cellular outcomes of the perturbations and improve the predictions. The use of variance to capture the difference in outcomes is interesting, however, the use of P\*, the ideal distribution, is not clearly explained. Further explanation is required as to what the ideal distribution is.

As for the scoring function, it is a dot product of two terms. The first term is Vi, is the variance across the distribution, and therefore leads back to the missing explanation in the statistic. The second term utilizes partly the log of 2 \* the dot product of Qi(j) divided by the sum of Qi and Qj. It is not clear whether this equation is interesting or not, because the logic and process behind it is not clearly explained. I suggest that the author defines all the variables inside the equation and the reasoning behind it.

The bonus part of this solution suggest that we choose the target Q based on Time-series experiments. However, and as the author indicated, that will reveal to be a difficult and costly exercise.

Note: There were no references / citations given in this proposal.

**Review 5:**

Submission ID: 676657

The authors propose a statistic that represents the change from unperturbed to perturbed cells obtained from screening . However, the authors failed to mention that this formula is the Euclidean distance formula in higher dimension. Thereafter, the authors state that the rate of growth G is reflected by the cycling state and assume that a low cycling state probability indicates a low growth rate. They based this assumption on a spearman correlation score of 0.47 and a p-value of less than 0.05. However, correlations and p-values are an indication of a relationship between the variables and not to be confused with causality. This score and p-value doesn’t tell us that one value is driving the other.

The final proposed formula for the scoring function is rather sophisticated and it is not clear if it can produce better results than the actual formula. We need to understand the reasoning behind adding

Finally for the proposal for Q, the authors give an example of an ideal desired Q which is [0.6, 0.1, 0, 0.3, 0]. It is not clear how they arrived at these proportions or if there is any scientific method to come up or support this approach.

Note: There were no references / citations given in this proposal.

**Review 6:**

Submission ID: 676673

The author proposes to utilize the Kullbeck-Leibler divergence between two Dirichlet distributions parameterized by the normalized compositions PQ and Q as a statistic. While the approach is valid, there is no sufficient development of the thought process or reasoning or even the equation itself.

Utilizing the standard errors of the above parameters estimate to propose gaussian approximations of the parameter posteriors, sample from those posteriors, and evaluate the distribution over KL divergences is a sound approach. However, this approach needs further development and explanation and implementation.

# 676620

## Strengths

* Very clear write-up. All of the proposed elements of the score are clearly motivated. Computationally feasible.
* The distance metric is appropriate for the task at hand. The metric gives a very nice and interpretable measure on how much the cell state distribution diverges from the unperturbed distribution. Confidence on the observation based on total cell numbers is also directly measured by the proposed score.
* Although no new statistic to measure the cell state is proposed, they give a good argumentation on why decision boundaries should not be used.
* They provide an interactive tool to visually inspect the effect of perturbations.

## Weaknesses

* Proposes a new scoring metric but not a new statistic.
* Total cell counts as a reflection of a beneficial perturbation effect are not considered in the score.
* The metric does not accommodate for the use of weights that give more importance to the change in a particular cell type.
* Noise/variability on unperturbed cell distribution is not considered.
* When proposing the incorporation of a LCB to take into account uncertainty generated by low cell counts, the correlation between cell state distribution and total cell counts is not considered nor discussed. Given that in this case uncertainty is not distinguishable from the biological signal we aim to predict, a discussion on how would this impact the score becomes necessary.
* In 4, it is not clear how this issue could impact the evaluation of the effects of a perturbation, where we are not aiming to predict cell states for unobserved perturbations with a new ML model, but simply measuring divergence from the desired cell state (defined by the same algorithm).

# 676651

## Strengths

* Both a statistic and a score -that are computationally feasible- are proposed in a

clear write-up.

* Measures if the cell remains functional by requiring a small distance to unperturbed cell gene expression. This distance accounts for variability in the observed control gene expression.
* The choice of distance/similarity metrics is appropriate for the task at hand.

## Weaknesses

* Lacks a bit of motivation in the score proposal.
* It may be more insightful to calculate shift in gene expression relative to cell type. It is not entirely clear what is the mean of unperturbed cells gene expression reflective of, when different cell types are averaged.
* Unperturbed cell state distribution is not taken into account.
* The metric does not accommodate for the use of weights that give more importance to the change in a particular cell type.
* No consideration of total cell counts and its effect, neither as a measure of observation uncertainty neither as a biological signal of interest.
* Cell type classification boundaries are not considered.

# 676630

## Strengths

* Computationally feasible and flexible proposal with a clear write-up.
* Measures deviation from unperturbed cell state.
* Allows to give more importance to cell types of interest by weighting.
* Confidence on the observation based on total cell numbers is directly measured by the proposed score.

## Weaknesses

* Proposes a new scoring metric but not a new statistic.
* Lacks a bit of motivation in the proposal.
* Absolute distance may not be the most appropriate distance metric for this case.
* Total cell counts as a reflection of a beneficial perturbation effect are not considered in the score.
* Noise/variability on unperturbed cell distribution is not considered.
* When penalizing low number of total cells, the correlation between cell state distribution and total cell counts is not considered nor discussed. Given that in this case uncertainty is not distinguishable from the biological signal we aim to predict, a discussion on how would this impact the score becomes necessary.
* Cell type classification boundaries are not considered.

# 676644

## Strengths

* Both a statistic and a score -that are computationally feasible- are proposed in a clear write-up and with appropriate motivation.
* The proposed statistic takes into account perturbation effects, by computing the distance to the unperturbed gene expression.
* Present an interesting approach to measure confidence in cell type labels and consider that in the actual score.
* Accounts for uncertainty in perturbations with very low cell numbers with a

thresholding approach so that, once the confidence is high enough, higher weight is not given to perturbations with more cells.

## Weaknesses

* It is assumed that a general shift in gene expression is desirable. It is not clear that this is true: why would we desire a change in the expression of houskeeping genes? A perturbation that only impacts the expression of a small set of genes may be more precise.
* It may be more insightful to calculate shift in gene expression relative to cell type. It is not entirely clear what is the mean of unperturbed cells gene expression reflective of, when different cell types are averaged. The variability of gene expression in control cells is also not accounted for.
* When scoring the effect on cell states distribution, it may be interesting to use the control cell distribution to measure divergence from baseline distribution.
* The metric does not directly accommodate for the use of weights that give more importance to the change in a particular cell type.
* Total cell counts as a reflection of a beneficial perturbation effect are not considered in the score.

# 676635

## Strengths

* A flexible and simple score is proposed and clearly motivated. Computationally feasible.
* The proposal of using Euclidean Distance instead of L1 distance to penalize less differences spread across dimensions nicely fits the challenge objective.

## Weaknesses

* Proposes a new scoring metric but not a new statistic.
* Expected cell state distribution based on unperturbed cells is not accounted.
* No consideration of total cell counts and its effect, neither as a measure of observation uncertainty neither as a biological signal of interest.
* The metric does not accommodate for the use of weights that give more importance to the change in a particular cell type.
* Cell type classification boundaries are not considered.

# 675876

Challenge requirements (providing a statistic and a score) are not fulfilled.

**Strengths**

• Authors provide an interesting biological rational in how to select Q.

## Weaknesses

* No references are given for the biological rational introduced.
* The fact that progenitor exhausted cells respond better than terminal exhausted cells does not mean that exhausted cells should respond better than nonexhausted cells. Therefore, it is not clear why one would aim to expand the exhausted population.
* No new statistic provided, scoring function is for scoring a prediction of perturbation effects, not a perturbation effect.

# 676612

Challenge requirements (providing a statistic and a score) are not fulfilled.

Explains how to predict a gene cell state for new perturbations, but does not propose any statistic to summarize the effect of seen perturbations or a scoring function on how to evaluate divergence from desired effect.

675876

The authors propose a “statistics function” that measures a deviation between a predicted distribution over 5 cell states.

Strengths:

* The reference distribution P0 is used in the metric.
* The authors discuss numerical edge cases.

Weaknesses:

* The presentation of the proposed method is a weakness. It is not very clear which paragraphs were copied over from the problem definition in the challenge. Parts of the proposed method are hard to follow and very brief.
* The notation in the statistics function is somewhat unclear. What looks like L1 norms are used but those seems to contradict what is said in the text, which is a ratio of “state proportions”. **S(X) = | X | / | (Y - Y1) |** seems to be **S(X) = 1 / | (Y - Y1) |** if || indicates L1. In the case that **S(X) = | X / (Y - Y1) |** was meant here, it is not directly motivated how this would measure the desired phenomenon.
* It is not clear if the number of cells per knockout is used in the metric.

676611

The authors propose a statistic based on Kernel-PCA for part 1 and a class-sized weighted mean squared error for part 2.

Strengths:

* The authors motivate their solution approach for each challenge.
* In part 1, the authors use a projection of the gene-dimensional data to do comparisons between states.
* In part 2, the authors adapt mean squared (MSE) error to account for class imbalance, thus weighting the MSE statistic by some notion of certainty.

Weaknesses:

* The motivation for the kernel PCA score is hard to follow. It is not documented why **Q - Kernel\_PCA\_score (Pi)** is a meaningful deviation measure to the target distribution over discrete cell states.
* In section 2, the metric is apparently a sum over genes so it is unclear how this will be used to rank genes.

676614

The authors propose to use the expected composition of a perturbation obtained from a generative model to then compute. The authors mention usage of “pseudo-knouts”, which are cells with low expression of a target gene, which could be used to further inform a metric.

Strengths:

* The authors work through examples of their score function on the training data.
* The authors propose a statistic that can deal with output from generative models.

Weaknesses:

* The proposed metric only uses the mean of the predicted distribution of a generative model and does not account for uncertainty in the data sample or the prediction itself.
* The discussion of pseudo-knockouts is hard to follow.

676628

The authors discuss a metric that is derived from a deviation in a gene-expression embedding space and is also predictable by a sequence model of the sequence predictors of the gene, such as TSS-proximal nucleotide sequences.

Strengths:

* The authors propose the usage of gene sequence features to constrain a score.
* The authors propose to measure distances in a gene expression embedding space, thus harnessing a low-dimensional representation of the data.
* The authors show some initial results based on their metric.

Weaknesses:

* While the constraint imposed by a model using gene sequence is interesting, it is not clearly motivated why this would be beneficial and why the sequence constraint is valid.
* The method involves non-convex optimization problems which may be detrimental to usage in practice and are not discussed from that angle in the solution.

676646

The authors motivate a few key characteristics of the data, including distributional uncertainty and its dependency on total cell count per perturbation, and the meaningfulness of discrete cluster boundaries. They then propose means to tackle these difficulties and propose a score function that is defined based on a probabilistic model of compositional data.

Strengths:

* Great presentation of motivation and method, including analysis figures and code.
* Inclusion of total cell count by perturbation in the metric and a discussion of its relation to distributional uncertainty when regarding the distribution of cells over the 5 states.
* Soft cell state definitions using a classifier in a gene expression embedding. This may alleviate artefacts that arise from misclassifications at the state boundaries.
* Proposition of a probabilistic model of the compositional observation of the cell state distribution that can be used to score a perturbation against the reference distribution.

Weaknesses:

* Section 3 would have benefited from more detailed formulas.

676666

The authors define a metric based on L1 distances between compositions and the total number of cells in the perturbed condition.

Strengths:

* The authors include the total number of cells per perturbation in the score.
* The authors define the metric as a sum of two distances, both from the reference condition and from the target condition, thus constraining the metric to reflect deviation on a data manifold more closely and potentially stabilizing the metric to outliers.

Weaknesses:

* The weighting of the different terms in the metric could be discussed.
* The relation of the proposed approach of measuring distances to two reference points could have been contrasted with distance measures on the data manifold.

676669

The authors propose the usage of gene ontology information to a deviation measure.

Strengths:

* The authors propose the usage of information from the GO database to relate individual genes to each other.
* The authors discuss weaknesses of the proposed usage of GO.

Weaknesses:

* The proposed metric is not clearly defined.
* The total number of cells per knockout does not seem to be used in the metric.

# 676583

The proposal explains a summarisation strategy to summarise the gene expression by reducing the dimensions. It proposes two scoring functions for perturbation and one for cell states. The proposal exhibits a limited understanding of variances. It appropriately incorporates novel features and is to an extent computationally feasible.

# 676614

The proposal by itself does not mention a statistic but proposes two scoring functions for each of the two types of therapies. The scoring function uses scaling factors which are pseudoknock out expectations calculated from the unperturbed cell data. This is a novel proposal and the proposal is computationally feasible. The scoring function is clearly explained and motivated. **676630**

This submission does not provide a statistic but proposes a measurable and computationally feasible scoring function. The submission proposes that the scoring function is calculated from weighted sums of L1 losses but fails to provide a valid methodology for choosing the weights. Since the calculation merely uses ratio between weighted L1 losses, novelty is limited and exhibits an un-compelling relationship between variance and the sample size.

# 676651

The proposal clearly explains both a statistic and a scoring function. Though the statistic is the same as that provided in challenge 2, the scoring function is novel. The motivation behind choosing the scoring function is also well explained. It proposes the distance function as the square root of inverse of a covariance matrix multiplied by vectors, the computation is well established and the score is a scalar and hence computational feasibility is proper.

# 676654

The proposal clearly explains and motivates both a valid statistic and a scoring function.The proposed statistic and scoring function incorporates novel features such as the variance and taking the logs. The function provided can be calculated from the desired Q and the variances.

# 676658

The submission compellingly explains both the statistic and the scoring function. The proposal is supported by compelling reasoning. Both the statistic and the scoring function is documented in a highly novel manner. The submission also provides intuition into how the scores can be calculated accurately.

# 676661

The proposal only provide a scoring function which is the KL-divergence and does not mention a statistic. Explanation for KL-divergence is provided but does not indicate why it is an appropriate choice. KL-divergence is a novel way of computing the distance between P and Q. The submission does not provide any explanation on what the metric P stands for and thus how it is calculated.

# 676673

Strengths:

* Considers the uncertainty in the score function by MLE
* Good motivation presented at the beginning. Weakness:
* No explanation for the statistic nor for the score function
* How KL between Pq & Q (What is Pq??)
* No mathematical expressions
* No explanation on the process (Ranking process)

# 674631

Strengths:

* Clearly mentioned the s(.) and scoring function
* The mathematical expression makes sense. Weakness:
* The uncertainty element from the distribution could have been explained more.
* L1 loss was used in the previous challenges as well. Hence, innovation could have been more.

# 676620

Strengths:

* Good motivation presented.
* Good formulation & explanation given.

Visualization acts as a bonus.

Weakness:

* The formulation & logic behind LCB as an uncertainty modification is a complicated addition in the metrics.
* I have my concerns with absolute distance between vectors. The TVD choice for vector distance could have been something more evolved like Euclidean distance as it captures the moments more accurately.

# 676611

Strengths:

* Novel technique used where the author talks about an alternate attempt at summarizing gene expression vector
* Good attempt in explaining the chronology of concepts applied.

Weakness:

* I am skeptical about the part of formula where he direct does this operation: **w2 \* abs (Q - Kernel\_PCA\_score (Pi))**

This intuitively doesn’t make much sense to me. [even if

Kernel\_PCA\_score (Pi) is a 5-dim vector]

* Is kernel\_PCA\_score(Pi) normalized that sums upto 1?

# 676670

Strengths:

* Good presentation & extra points for the block diagram describing the methodology.

Novel technique of getting the statistic which is not based on cell state proportion vector.

Weakness:

* Absence of any mathematical expression. It would have helped especially for the case of scoring function.
* Not clear how the scoring function works. The explanation could have been more transparent

# 676614

Strengths:

* Good approach for presenting 2 scoring functions, one for each kind of therapy.
* Good distinction between scoring function, the statistic and how the scoring function uses the statistic.
* The approach is simple to understand yet I feel it’s a good metric proposal.

Weakness:

* I found the novelty factor a bit missing. It seems it is highly inspired from the metric proposed in Challenge 1 & Challenge 2.
* It seems for the Immune checkpoint blockade therapy the author didn’t consider the factor of minimum number of cycling state.

# 676644

Strengths:

Good motivation to the metric problem & explanation of all the related terms.

* Good starting point to begin with calculating the submetrics followed by the main metric.
* The final scoring function is the sum-squared of both the statistic’s distributions: the gene expression & the cell state vector. This is a good novel approach of taking into both the variants.
* It was simple to understand metric and yet it seems quite effective.

Weakness:

* Choosing the sample size as a hard 30 has my concerns since it might not be relevant to this domain.
* I think it will be difficult to compute the metric programmatically even form time-complexity point of view.

**Summary of comments**

Two submissions were not valid. Three proposed to use an estimate of the observed cell state proportions as the statistic, with appropriately weighting to construct a scoring function. The other two used predictions of cell state proportions derived from some property of the gene (without describing the observed gene expression data).

**676619**

Authors propose to learn a GRN on the unperturbed data that estimates the proportions and use its predictions on the perturbed data as a measure of distance to target

* + Provide a strategy for estimating proportions/scoring function
* - No results illustrating proof of concept
* - The statistic is just the L1 difference between predicted proportion and desired one; it doesn’t actually summarize an observed distribution since it is based on prediction from unperturbed.
* - Variable cell numbers not relevant

**676622**

Not a valid submission for this challenge. No statistic/scoring function proposed; unclear writeup.

**676630**

Author proposes to score perturbations by the ratio of distance to target proportions in the perturbed state to distance to target proportions in the unperturbed state. A weighting term is proposed to weight states in the distance computation, but no suggestion for how to choose it, limiting feasibility.

* + Weighting term that scales by number of observed cells to account for confidence.
* - Some implementation to score the perturbations.
* - No new statistic proposed (i.e., the statistic is the observed cell proportion vectors).

**676634**

Not a valid submission. A jupyter notebook is provided but no writeup. I had a look through notebook, but it seems to be some prediction task – no mention of proposed statistic or scoring function.

**676643**

Authors estimate cell state proportion vector as a softmax weighted average of the number of cell-state specific gene sets (computed via GSEA) that the perturbed gene appears in. Next, they propose using cosine similarity to score distance to target. Seems like an ad-hoc way to determine gene

* + It makes sense to use gene ontology to try and estimate perturbation effects
* - Arbitrary correspondence (softmax( s/10) - unclear why this statistic should correspond to observed vectors.
* - Statistic does not summarize an observed gene expression distribution. No results demonstrating their statistic corresponds to observed proportions (unclear why they should). Unclear how reliable the GSEA pathways they have observed are since there appears to be a lot of redundancy and.
* A more direct GRN inference method would have easier to validate.

**676657**

Authors propose to use 1 - |a+b-c| as the statistic, where a,b,c are proportions of prog/eff/exhausted.

* + Authors propose to score perturbations by reweighting to penalize those without cycling cells (seems like a good idea).
* - Unclear how valid their parameters are – an arbitrary weighting term is not computationally generalizable.
* - No proposal for how statistic should be estimated

**676670**

Authors propose to use cluster quality as the statistic.

* + Interesting idea to think about cluster quality as a requirement for good perturbations.
* - The statistic wouldn’t distinguish 100% cells all exhausted with perfect clustering or 100% cells all progenitor with perfect clustering, so would not be usable downstream – but that’s not to say some variant of it couldn’t be applied.
* - Scoring function requires knowledge of proportions.

# Submission: 676661

**1.1 Summary**

This submission proposes three incompletely described statistics, and no scoring function.

**1.2 Does the proposal provide a statistic and a scoring function, and are they both clearly explained and motivated?**

The first statistic suggested measures the Kullback–Leibler Divergence (KLD) between ‘distributions’ for T cells in a perturbed versus unperturbed condition. The probability distributions between which the distance is measured is not specified (it is presumed to be *Q* - the distribution over final T-cell states, as described in Challenges 1-2 - but the notation is not clear). A normalization process is further mentioned, but not specified. The second statistic is described as a ‘graph neural network based metric’ - but the way in which the graph is composed in not well specified (“construct a graph consisting of gene nodes to describe the joint probability distribution of genes...each cell sample as a graph...”) making it difficult to understand how it’s intended to be used. The third metric - a ‘dynamical system based metric’ is not specified at all, beyond it’s name.

**1.3 Does the proposed statistic and scoring function incorporate novel features beyond what was used in challenge 2?**

The author suggests some concepts / tools that may be fruitful - for example, it may be possible to capture information about the joint probability distribution / shared information for each pair of genes, and use that to make better predictions about as-yet-unseen perturbations - but these are not fleshed out, into a statistic or a scoring function that can be put to use as-is.

**1.4 Is the proposal computationally feasible?**

Yes, to the degree that it’s described.

**1.5 Conclusion**

This proposal suggests some interesting avenues but does not satisfy the original ask for a novel metric, and a scoring function by which that metric can be used to rank novel perturbations.

# Submission: 676644

**2.1 Summary**

This submission proposes two metrics - one measuring the degree to which the perturbation changes gene expression relative to the unperturbed state (*SiP*), and the other measuring the degree to which the distribution over resulting cell states match the desired distribution (*SiR*) - as well as a method by which to combine these into a single final score that can be used for ranking. The final metric values *both* a greater degree of differentiation between the perturbed and unperturbed gene expression profiles, *and* the similarity between the state distribution achieved for the perturbed condition relative to the desired state distribution.

**2.2 Does the proposal provide a statistic and a scoring function, and are they both clearly explained and motivated?**

The proposal provides for both a statistic (in fact, two), and a scoring function, as required. The derivation of each of the two statistics is relatively well described, though the reader is left with some questions.[[1]](#footnote-1)**The motivation for the first statistic though - that which measures the distance between the gene expression profile for the perturbed cells and the gene expression profile for the unperturbed cells - is unclear:** If the objective is to rank unseen perturbations relative to their ability to achieve a desired distribution of cell states - the information provided for this statistic is immaterial. When combined with the second statistic - which measures the similarity between the achieved and the desired distribution of cell states - the first statistic serves to prioritize those perturbations which which achieve the same end state, by having made a greater impact to the gene expression distribution, but the proposal never describes why this is desirable (achieving the same ends, via different means). The submission / idea described here would still work, and would be more parsimonious, if this first statistic were simply dropped.[[2]](#footnote-2)

**2.3 Does the proposed statistic and scoring function incorporate novel features beyond what was used in challenge 2?**

The second statistic (*SiR*) - measuring the degree to which the distribution over resulting cell states match the desired distribution - provides for a measure of confidence in cell state attribution, which is useful (in the calculation of the distribution of cell states, cells that sit closer to the centroid for a given state are assigned more weight than those which sit farther away). For the initial statistic, there is also a novel initial dimension reduction step - but the value added here (versus the method used in Challenges 1 and 2) is never explicated. The initial statistic also attempts to account for the size of the sample used to estimate the gene expression distribution (which is admirable), such that perturbations that result in a greater number of cells over which the expression can be estimated are assigned a greater ‘weight’ (which results in a greater ‘distance from unperturbed’ value). In so doing, however, the authors state that ‘...a sample size of 30 is usually considered large enough in statistics...’ and then go on to use 30 as a magic number in their calculation. Whereas this will move them at least in the right direction (properly assigning more weight to perturbations that result in more than 30 samples, relative to those that result in fewer) the use of a static number here (especially without some idea of the desired effect size and a power analysis) is bad, and reduces the positive impact of having included this weight. A simpler, and better, solution would have been to simple weight by how many samples (cells) were produced.

**2.4 Is the proposal computationally feasible?**

Yes.

**2.5 Conclusion**

This proposal is thoughtfully written - the reader can follow the concepts they’re trying to capture (even if the specific examples raise questions), and the way in which they’re trying to use them. It is less well motivated: It’s unclear, for example, why the first statistic is valuable; it would appear to detract from, rather than add to, the final ranking that the scoring function would produce.

# Submission: 675654

**3.1 Summary**

This proposal begins to suggest a competitive method for identifying the best genes - with competing experiments run by ‘companies’ - but only barely.

**3.2 Does the proposal provide a statistic and a scoring function, and are they both clearly explained and motivated?**

This proposal includes neither a statistic nor a scoring function.

**3.3 Does the proposed statistic and scoring function incorporate novel features beyond what was used in challenge 2?**

This is not applicable, given that there are no statistics or functions.

**3.4 Is the proposal computationally feasible?**

No.

**3.5 Conclusion**

This proposal is incomplete and cannot be evaluated as such.

# Submission: 676666

**4.1 Summary**

This submission is similar in concept to [676644]; it proposes a final score that values both a greater distance between the perturbed state distribution and the unperturbed state distribution, *and* a lesser distance between the perturbed state distribution and the desired state distribution. (Put another way, ‘...the further the state vector of a condition *Qi* from unperturbed condition *Q*0, and the closer the state vector from the desired *Q* the better.’)

**4.2 Does the proposal provide a statistic and a scoring function, and are they both clearly explained and motivated?**

Yes. The proposal preserves the statistic *Q* utilized in Challenges 1 and 2 (a distribution over 5 cell states), and then provides for a novel scoring function using this statistic. (The way in which *Q* is derived for a novel perturbation is not specified, but the reader presumes that it could be [any of the methods proposed in Challenges 1 and 2]. This could be more clearly specified, however.) The motivation for keeping the default statistic - versus any of the variants proposed in other submissions - is not described; this would have been nice to include in addition. Finally - and similarly to notes above for Submission 676644 - the motivation for including in the final score a weight for the degree to which the perturbation is different than the unperturbed state, when we’re already including a weight for how similar the perturbation is to the *desired* state is unclear, and never addressed by the authors.[[3]](#footnote-3) The score would be more parsimonious without this weight; if it adds important information, that should be described in the proposal.

**4.3 Does the proposed statistic and scoring function incorporate novel features beyond what was used in challenge 2?**

The scoring function includes a term that lends greater value to perturbations that result in a larger number of cells from which to estimate the statistic - which is generally desirable. In this case though, the term is defined relative to the sample size for the unperturbed state (rather than in absolute terms, or relative to some *a priori* estimate of variance). This is ok *if we can presume that the sample size for the unperturbed cells is sufficiently large*. If that is not a given, however, this metric will not achieve that for which they’re going here (i.e. the relative proportion alone tells you nothing; it is only meaningful if the denominator is sufficiently large in absolute value to allow you to confidently estimate variance for that unperturbed population[[4]](#footnote-4).)

**4.4 Is the proposal computationally feasible?**

Yes.

**4.5 Conclusion**

This proposal is simple, and easy to understand - but like Submission 676644, the reader is left with some questions about some terms, and - more importantly - there’s a key factor within the scoring function that’s never justified or motivated (the degree to which the perturbed state differs from the unperturbed state, notwithstanding the degree to which it matches the desired state).

# Submission: 676669

**5.1 Summary**

This proposal appears to suggest the general concept of looking at gene “similarity” (of a test case to a known case), when estimating the final state distribution of an unseen perturbation. It describes a method based on the RNA-seq data provided in Challenge 1 - as well a method based on third-party ontologies (e.g. the [Gene Ontology Resource)](http://geneontology.org/), though the latter is only obliquely described.

**5.2 Does the proposal provide a statistic and a scoring function, and are they both clearly explained and motivated?**

Neither are clearly described. We infer that they are suggesting a statistic that captures the similarity of a test gene to a known gene, using the RNA-seq data for the unperturbed population. That is: If you have a gene *i* for which you want to estimate the final cell state distribution *Qi*, you can do this by finding - from amongst the genes that HAVE been perturbed, and for which we have validation data (*N*=66), that which is most similar, in terms of it’s unperturbed expression profile. That known gene’s cell-state distribution, when perturbed, is then used as the estimate for the final state distribution for a sample in which gene *i* is perturbed. Even if this inference is correct, however, no scoring function is provided, and it’s thus impossible to rank our unseen perturbations in terms of the degree to which they match [any given desired cell state distribution.] One could presume they may do so using the standard loss function used in Challenges 1 and 2 - but again, this is not explicated.

**5.3 Does the proposed statistic and scoring function incorporate novel features beyond what was used in challenge 2?**

No.

**5.4 Is the proposal computationally feasible?**

Yes, insofar as our inference above (as to their intentions) is correct.

**5.5 Conclusion**

The proposal starts to suggest at some novel concepts - taking unperturbed expression similarity and/or similarity in a third party ontological space into account - but it is never drawn out to the degree that it could actually be implemented by anyone reading this proposal. Concrete examples would be helpful here.

# Submission: 676333

**6.1 Summary**

**6.2 Does the proposal provide a statistic and a scoring function, and are they both clearly explained and motivated?**

Yes, both a statistic and a scoring function are provided. The proposed statistic (*Pi*) is the distribution of gene expression in the full/high-dimensional (*N*=15,077) space. The proposed scoring function includes a term that compares the gene expression distribution for the novel perturbation(*Pi*) to that for the unperturbed population (*P*0), a term that compares the gene expression distribution for the novel perturbation(*Pi*) to the expression distribution associated with the desired final state distribution (*Q*), a term for growth rate, a term for sample size, and a ‘penalty’ term, that attempts to capture the degree of confidence we have in applying the final cell state labels (that would go into *Qi*). There are some terms within this scoring function that are not well defined (such that the reader could implement them him/herself): First, it’s not specified how the estimate of the final state distribution is related to (*Pi*), the estimated gene expression distribution. In order to implement this scoring function, one would need to know how to obtain an estimated gene expression distribution (*P*) given only the desired final cell state distribution (*Q*) - given that the proposal tries to calculate the KLD between this idea *P* and the *Pi* for a given perturbation. Second - like Submissions 676644 and 676666 - the final score lends weight to perturbations that get closer to the final desired cell state distribution *via introducing changes in the gene expression distribution*, but the reason why this is desirable (over and above simply getting closer to the final desired cell state distribution, or the gene expression profile that represents that) is never addressed directly. Given that the score is simpler without this extra term, its inclusion should explicitly motivated. Finally, the authors don’t address how the growth rate or penalty terms in their scoring function are calculated. The penalty term is described as ‘...a measure of the deviation from the classification boundaries, such as the Mahalanobis distance from the closest boundary...’ but it’s never specified on which population this is calculated, or how.[[5]](#footnote-5) It also appears as if some of the signage in the scoring function is backwards (e.g. as written, the score would appear to favor perturbations that are farther from the desired state distribution, and whose statistics were estimated from a smaller sample size).

**6.3 Does the proposed statistic and scoring function incorporate novel features beyond what was used in challenge 2?**

The statistic does not; but the scoring function does (insofar as it attempts to capture the impacts of growth rate, sample size, and the confidence that we should have in the classification of the final states).

**6.4 Is the proposal computationally feasible?**

This is unclear - the authors never describe how the statistic *Pi* would be estimated for novel perturbations. It is suggested that an ML-based approach - similar to what they used in Challenge 1 - could be used, but given the high dimensionality of the proposed statistic, that would require a substantially larger training set than was provided, to do this confidently, if it were possible at all.

**6.5 Conclusion**

Whereas the general concepts that the proposal touches on make sense, and provide useful information, this proposal raises questions it does not answer, may be computationally infeasible, and does not give enough information with respect to some of its terms that the reader could confidently implement it him or herself.

# Submission: 676611

**7.1 Summary**

This proposal attempts to capture both the distance between the gene expression distribution for the perturbed sample and that for the unperturbed sample, and the distance between the gene expression distribution for the perturbed sample and the distribution representing a sample that would result in the desired distribution of final cell states, and weights both of these in its final score (with adjustment based on the size of the sample that falls within each final cell state).

**7.2 Does the proposal provide a statistic and a scoring function, and are they both clearly explained and motivated?**

Both a statistic and a scoring function are provided, but they are not motivated. The primary ‘novel’ attribute here is the use of kernel PCA (vs. linear PCA, as in Challenges 1 and 2). In the introductory videos for the challenge, the instructor noted clearly that the final clusters in UMAP space matched up with the experimenters’ *a priori* knowledge (with cells that had high expression in genes known to be associated with certain final states falling cleanly within the expected clusters). If linear PCA and then UMAP reduction provided bounds that clinicians agreed with, it is unclear why the kernel approach is motivated - and this is never addressed explicitly by the authors.

**7.3 Does the proposed statistic and scoring function incorporate novel features beyond what was used in challenge 2?**

The proposal attempts to provide more weight to estimates derived from larger samples via the addition of Weight Mean Square Error (WMSE) vs. MSE.

**7.4 Is the proposal computationally feasible?**

Yes.

**7.5 Conclusion**

This proposal is not clearly motivated, and as such is difficult to score highly.

1. For example, the calculation of *SiP* includes an initial dimensionality-reduction step wherein the authors calculate the correlation coefficients for each pair of genes (*N*=15,077). For ‘highly correlated genes’ - defined as those with correlation coefficients *>* that which represents the 80th quantile of all the correlation coefficients calculated - they keep the gene with larger variance. The authors state that they end up with 8,334 genes using this method - but this isn’t possible; the value should be higher. Either this number or the cutoff used is incorrect. Nonetheless, the concept makes sense - it’s just the concrete example value that is confusing, and could be further elucidated. [↑](#footnote-ref-1)
2. The motivation for the novel form of initial dimension reduction used in the calculation of the first statistic is also never described. In Challenges 1 and 2, initial dimension reduction is achieved by down-selecting genes to those with the greatest variance across samples. Here, that is replaced with a pair-wise comparison between all genes, with the authors selecting, from the mostly highly correlated pairs, the gene with the greater variance. The reader can guess as to why this is a preferable route, but it’s never explicitly addressed - which is important, given that the proposed method is more computationally complex than the original. [↑](#footnote-ref-2)
3. Submission 676644 measured the degree to which the perturbed and unperturbed populations were different within expression-space, whereas Submission 676666 measures this within cell-state-space, but the concepts are the same. [↑](#footnote-ref-3)
4. This also presumes that the variance is similar within the perturbed and unperturbed samples, which is not a given. [↑](#footnote-ref-4)
5. Are the classification boundaries here those that were identified within UMAP space for the perturbed data taken in isolation, or is there some implicit comparison to the classification space for the unperturbed data? How is the ‘closest’ boundary identified? If the ‘penalty’ is estimated as the Mahalanobis distance from the closest boundary - versus from the centroid of a cluster - does that not penalize clusters that are both compact and distict from each other, and lend extra weight to clusters that are diffuse and overlapping? This would be exactly the opposite of what we would want, if we were trying to capture classification ’confidence.‘ [↑](#footnote-ref-5)