Given my particular approach to classifying the results of perturbations in parts 1 and 2 via gene ontology (though the actual GO algorithm is only visible in the submission for Challenge 2), I strongly believe that 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*. More particularly, while keeping the total number of cells resulting from the knockout high is a positive contributor to a gene target worth knocking out, I believe that the gene ontology approach is imperfect.

The dataset that we used to form our Gene Ontology tree was incomplete in the sense that many of the Uniprot IDs (generated using a general expression algorithm on the FASTA file summarizing the *mus musculus* proteins available on the Uniprot website) were not represented in the gene ontology tree. Furthermore, many of the “common name” genes were not even present in the Uniprot dictionary, which meant that many genes were lost along the way. Adding the number of GO affiliations as a positive contributor to the score of anticipated knockouts could help account for the bias introduced by better-explored regions of the GO tree.

In the alternative approach that we used for Challenge 1, where we looked at the unperturbed RNA-seq data to find the closest correlating validated genes to the test genes of interest, we found that taking the mere mean of the top 5 best matches loses a lot of valuable proximity information. More particularly, not all test genes had really closely correlating validation genes from which to predict the cell state distributions. This is why I recommend instead implementing a weighted average to the results of the correlation. It should be noted that these results arise from the direct unperturbed cell-by-cell comparison of the progressive changes in each validated gene as compared to an individual test gene. For more please see our writeup for Challenge 1.