January 05, 2020

Dear Dr. Doerr and Members of the Nature Methods Editorial Board,

We are writing for a presubmission inquiry regarding the suitability of our manuscript, “Improving Deconvolution Methods in Biology through Open Innovation Competitions: an Application to the Connectivity Map,” for publication as Brief Communication in Nature Methods.

Deconvolution problems are ubiquitous in many areas of science. In the context of biomedical research, the problem often consists of acquiring gene expression levels from distinct populations (cell types, tissues, and genes) using random composite measurements (e.g., using a single analyte, or sensor, to measure signals from distinct cell types). Existing deconvolution approaches work well in many specific settings, but maybe suboptimal to the increasing availability of biological data. Machine learning techniques can potentially improve the current means to identify and isolate measurements from distinct populations because of the ability to capture automatically nonlinear patterns that are hard to model otherwise; especially in complex and massive datasets as those frequently used in biomedical research. However, entering these new machine learning techniques in the field presents several challenges. Some of which is validation, adaptation to specific datasets, and identification of the best machine-learning approaches to specific problems.

To address these challenges, we generated a novel experimental dataset with the transcriptional response of approximately 1,000 genes to 122 different perturbagens (shRNA and compounds) with several replicates for a total of over 2,200 gene expression experiments; and based on these data, we explored different machine learning approaches through an open innovation competition.

We show that (1) machine-learning approaches, such as Random Forests and Convolutional Networks, significantly improved performance compared to the benchmark, which was based on a k-means approach; (2) traditional methods such as gaussian-mixture models, however, can potentially outperform machine-learning approaches in prediction accuracy, as well as in computational speed; (3) the winning approach, which was based on a random forest, a popular machine-learning technique, achieved the highest global correlation with the ground-truth, the lowest inter-replicate variability, and, compared to the benchmark, was able to detect more than a thousand additional differentially-expressed genes, while improving the detection precision at the same time. This provides evidence of the tremendous potential of using random-forest approaches for deconvolution methods in biology.

We believe this work will bring much-needed insights into the field on how to conduct the experimentation of these machine learning methods, especially in settings where more traditional deconvolution approaches are the standard and can potentially represent “local-minima” solutions to the problem. We also believe that researchers will find the generated dataset a powerful research tool for benchmarking their methods, as well as a useful resource for multiple applications.

Thank you for your time and consideration; we look forward to your comments and feedback.

Yours Sincerely,

Andrea Blasco and **Karim R. Lakhani**

Harvard Business School

Aravind Subramanian

Broad Institute