PharmaHacks Challenge: Genomics

1. Preamble

If you haven't done so already, **please read our background guide**.

If you wish to perform feature generation using scFeatures, you will need to use R. However, to train ML models, you have the choice to keep using R or export the data and load it from Python. This may be wise, depending on your approach.

2. Problem statement

Your mission over these two days is **to perform patient outcome classification on COVID-19 patients (i.e., determine whether a patient will classify as mild/moderate or severe/critical based on their scRNA expression data)**. In order to complete the challenge, use the guidance of the benchmarking paper by Cao et al. [1] and use the dataset of Schulte-Schrepping et al. [2] linked here as a starting point (.rds file). We have also made a mirror[1] available in case of issues with Google Drive.

If you have pre-downloaded the `main_rds.7z` file previously, please do not download the data again, as this would put undue strain on the venue's network. In that case, you can now open `main_rds.7z` with the following password: `vACsgr4TmcsX01`.

2.1. Getting started

1. The benchmarking paper makes use of an R package (which was developed by the same group) called scFeatures to generate multiple types of molecular representation of each individual patient from single cells of individuals. The documentation of the scFeatures package for feature generation can be found here. Optionally, the scFeatures paper can be accessed here if you want to read more about features as a concept.

   You can download the scFeatures package in R using the command below:

   ```
   devtools::install_github("SydneyBioX/scFeatures")
   library(scFeatures)
   ```

2. Make sure you keep the data files on a path that you set as working directory. If you have trouble doing that, check the link here.
3. You need to access single-cell RNA sequencing data from the Seurat object. You can find normalized data and metadata by inspecting the object.
4. Make sure to run gc() command every once in a while to free up unused memory.

We will provide you with a way to submit your **code** and the **presentation** that you will show the jury should you be selected for the final round of judging.

2.2. Making a submission that stands out

- Think about the biology. Look at it more as an experiment but instead of pipettes, you work with codes and ideas. Having too many features can lead to overfitting, so you should try to minimize them. A biologically sound decision in feature selection can go a long way.
- Justify your choices. The "why" is as important as the model's metrics.
- Make use of information in the benchmarking paper [1] to make better decisions.

[1] http://5.135.137.166/pharmahacks/genomics/

- Incorporate more datasets from table 1 of the paper [1] to make your model more reliable. We prepared an object from a subset of Stephenson et al. [3] data that you can find here. This will be more convenient to work with than the original data.
- Your presentation could include graphs, F1 score, cross validation strategy, feature selection strategies, ML models used (and why), SHAP values, and report of features that significantly improved model performance, which could shed light on the biology of patient response to the COVID-19 outcome.
- Who knows, you may find a new potential drug target for COVID-19.

Good luck!

3. Getting help

If you are stuck or need clarification, feel free to ask questions and/or discuss with other participants on our Discord server.

4. Reproducibility

Make your code **clear and reproducible**. We will thoroughly evaluate the winning entries after the event and **disqualify post-hoc** any solution that is not reproducible with reasonable effort.

Your submitted solution may not rely on any data file that is not publicly available, unless you provide the **entirety** of the code required to create it from public data. It is okay to preprocess the data in any way you desire, but it is not okay to do so without explicitly stating the process by which you do so. Otherwise, we may not be able to verify that your model functions as you claim, and will be forced to disqualify your team.

Unnecessarily obscure code, with cryptic function/variable names and no comments, will be penalized.

Beyond preventing cheating, this measure is aligned with the principles of Open Science. It is more important than ever to ensure transparency in the scientific process. While making your solution clear and reproducible may seem like unnecessary overhead, it ensures that your work can be fairly evaluated (in this challenge) and that others can build upon it without duplicated effort (in general).

We will be on the lookout for any attempt at manipulating the results.

Bibliography

[1] Y. Cao, S. Ghazanfar, P. Yang, and J. Yang, “Benchmarking of analytical combinations for COVID-19 outcome prediction using single-cell RNA sequencing data,” *Briefings in Bioinformatics*, vol. 24, no. 3, p. bbad159, 2023, doi: 10.1093/bib/bbad159.

[2] J. Schulte-Schrepping *et al.*, “Severe COVID-19 Is Marked by a Dysregulated Myeloid Cell Compartment,” *Cell*, vol. 182, no. 6, pp. 1419–1440, Aug. 2020.

[3] E. Stephenson *et al.*, “Single-cell multi-omics analysis of the immune response in COVID-19,” *Nat Med*, vol. 27, no. 5, pp. 904–916, Apr. 2021.