Identification of potential treatments for COVID-19 using Regression Plane and machine learning

### Yiftach Savransky – 312141369, Amit Shakarchy – 313278889, [Colab link](https://colab.research.google.com/drive/173V5mT8U4Ku05RVzOvPDvbLtRKV_lOyx?usp=sharing)

# Introduction

SARS-CoV-2, a coronavirus that causes the contagious disease known as COVID-19, was initially identified in Wuhan, China in December of 2019. COVID-19 was declared a pandemic by the World Health Organization following its spread across the world. Currently, there are over 197M reported COVID-19 infections worldwide and over 4.2M reported deaths[[1](#r1)].

There is a need for fast identification of potential treatments for COVID-19 in human cells. We propose an approach that enables cellular image-based analysis in supervised manner with no domain knowledge required using the Regression Plane (RP) concept [[3](#r3)]. RP is a methodology for modelling and analyzing biological processes using 2-dimensional (2D) regression. Our approach is based on the ideas presented in the original paper by [[3](#r3)], however, we perform naïve modelling of the problem space. The goals of this study are to **identify potential treatments for COVID-19** and **analyze the effect of treatment’s dosage** on the treated cells. The suggested approach is used to reanalyze a published phenotypic drug discovery dataset to achieve these goals. An additional goal of this study is to examine the robustness of the proposed method by **analyzing how different modeling affects the biological conclusions** drawn.

Our study focuses on evaluating and applying our RP-based approach on the “RxRx19a” dataset, presented by [[2](#r2)]. The drug screening dataset, which aims to enable the evaluation of the effectiveness of treatments on active SARS-CoV-2 infection in human cells, consists of screening images and their corresponding deep learning embeddings. 1,670 compounds were tested in various dosages on two different cell types - normal human renal cortical epithelial cells (HRCE) and african green monkey kidney epithelial cells (Vero). Additionally, for the sake of comparison between the treated active cells to “healthy” cells, inactive SARS-CoV-2 ultraviolet-irradiated (UV-IR) and mock cells were included. Over 300,000 Immunofluorescence screening images (1024x1024) of the wells were taken in five channels, one for each unique stain. In addition, the authors provided embeddings (1024) of these images that were extracted using deep learning neural network trained for the identification of distinct phenotypic profiles.

As part of the paper’s authors analysis of the “RxRx19a” dataset they identified Remdesivir (as well as its parent nucleoside GS-441524) as the only antiviral tested with strong efficacy in suppressing the broad impacts of the SARS-CoV-2. Nonetheless, they suggested several compounds which demonstrated moderate to strong effects, by their analysis, to be further examined for treatment of COVID-19. Notebly, according to the authors’ analysis, Chloroquine and Hydroxychloroquine Sulfate showed no benefit. We aim to arrive at the same conclusions presented in this paper and enable efficacy analysis using our RP-based method which requires highly limited domain knowledge.

Szkalisity [[3](#r3)] proposed the **Regression Plane (RP)** methodology, a supervised, machine learning approach for performing image-based single-cell analysis of continuous biological processes using 2D plane modeling. 2D modeling of a biological processes enable the modeling of branching, circulating, parallel and crossing processes. In this paper, data was labeled by an expert with coordinates relative to the cell's state in the examined biological process, as believed by the expert. A regression model is trained to receive a sample and predict its position in the plane (that represents the predicted continuous cell state within the biological process). The trained model is then used to perform further analysis to derive biological insights.

There are two main limitations to the method presented in this paper, in regards to applying it to COVID-19 treatment discovery. First, the “classical” pre-processing steps of High Content Screening (HCS) are required. Typically, steps such as segmentation, illuminations corrections and hand-crafted feature extraction. Second, the results are highly dependent on the expert, both on the modelling of the Regression Plane and on the interpretation of the results. Our approach does not require major time-consuming pre-processing steps, and importantly does not require or depend on any significant domain-specific knowledge.

# Experiments:

## Data annotations & training process

Annotations of the dataset were determined by positioning the sample over the RP in a naïve manner. For each training sample (an embedding) we assigned 2D coordinates that correspond to one of 3 cell conditions- active, irradiated or mock. We selected 3 centroids over the RP to represent each condition. Gaussian noise is added to the coordinates to introduce a regularizing effect. Figure represents a possible mapping of the training samples to the RP. We then trained a regression model (multi-output Random Forest regressor) to receive a sample and predict its position in the plain. A full RP was created by projecting the predictions of the trained model on all the test samples to the 2D plane. Figure shows that the regression model has managed to distinguish between the different cell conditions (achieving score of 0.624 on the training set).

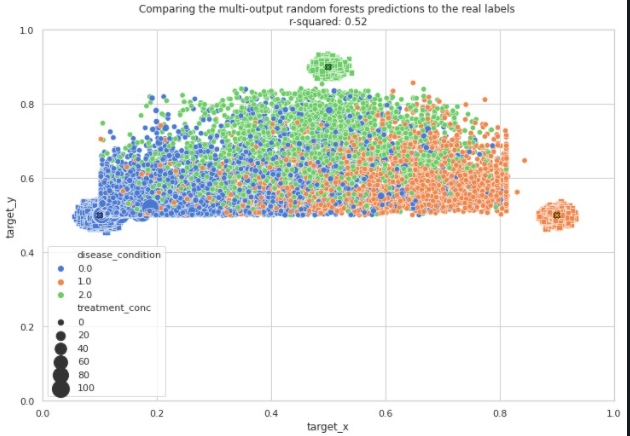


Figure 2: model's predictions over the regression plane

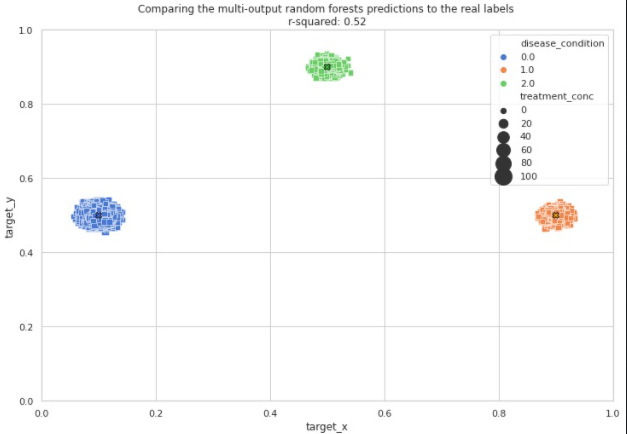
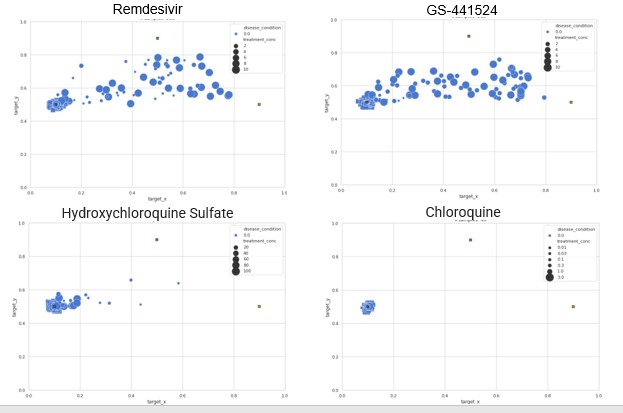


Figure 1: Regression Plane’s training set, according to our basic modelling.

## Projecting treatments on the Regression Plane

To examine whether projection on the RP reflects the efficacy of a single treatment, we project cells that were treated with treatments with a known efficacy level [[2](#r2)]. Figure 3 shows the projection of 4 examined treatments. Despite our lack of domain knowledge, the conclusions of Heiser [[2](#r2)] can be drawn from the visual projections. Both Remdesivir and GS-441524 demonstrate a large trajectory over the plane away from the active centroid (left), towards the mock centroid (right) and the irradiated centroid (top); neither hydroxychloroquine nor chloroquine’s trajectories spread towards the “healthy” centroids.

Figure 1: projection of specific treatment’s samples to the Regression Plane. The size of a single point represents the dosage of the examined compound.



## Measures for the Effectiveness of treatments

We tested several options for quantitative measures of the behavior of a treatment’s trajectory across the RP:

1. **Euclidian Distance** – the average distance of a specific treatment’s samples from the active condition centroid. A treatment’s efficacy increases as distance becomes longer. Similarly, a short distance to the “healthy” centroids is preferred.
2. **Cosine Similarity** – similarity of a specific treatment’s samples to the mock condition centroid. High similarity indicates on an effective treatment.

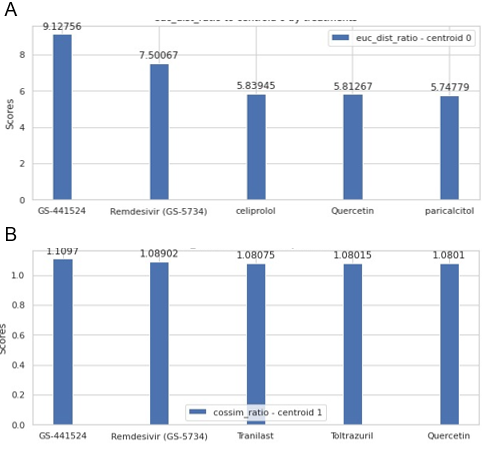


Figure 2: Efficacy of top 5 treatments tested. (**A**) Euclidian Distance. (**B**) Cosine Similarity

As can be seen in Figure 4, Remdesivir and GS-441524 hold the best scores in both measures. These results are in line with [[2](#r2)], thus reassuring the capacity of the formulated metrics. Additionally, our metrics can be used to identify new potential treatments. For instance, Quercetin was deemed as a compound with strong efficacy and therefore we would recommend it for further examination and trials.

## Modelling Drug Dosage Effect

The ability to predict the effect of drug dosage, may assist to further examine a potential compound (for dosage recommendation) or to identify new treatments that should be analyzed. Hence, we attempted to formulate a drug-specific function that receives coordinates on the 2D plane and predicts the dosage level required to reach these coordinates. We performed polynomial fitting of our dosage-dependent data to a 3rd order polynomial:   
 Our evaluations of goodness of fit, using chi square test and KL divergence test, were not statistically significant.

## Modelling the Problem Space

Biological insights derived using the RP methodology, are highly dependent on the modelling of the training set. We wished to investigate how the predicted efficacy of compounds alters for different modelling of the problem space. We focus our examination on plane modelling that may make the training process of the regression model harder. Figure 5 shows one such plane modelling, in which the active centroid is placed between the “healthy” centroids. The ability of the model to distinguish between the cell conditions reduces ( score of 0.174). In addition, the drugs that were found to be the most effective are different. However, paricalcitol that was found with the best efficacy by this modelling, is in fact a drug that is used for COVID-19 treatment. Remdesivir was identified as the 10th drug, despite the inferior performance of the model.

Figure 3: Euclidian distance of top 10 treatments identified using the challenging modelling.

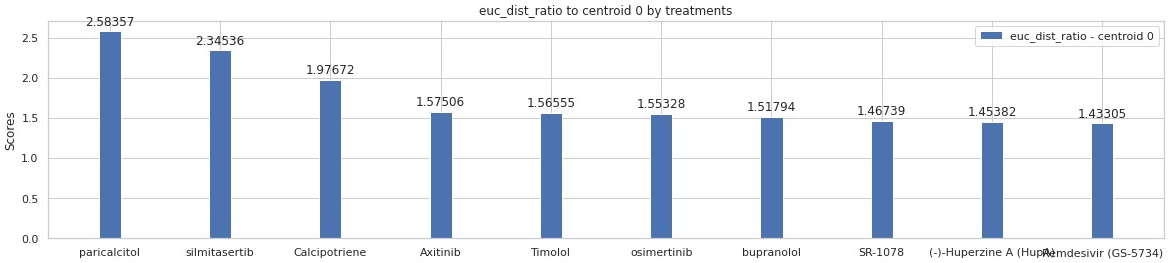
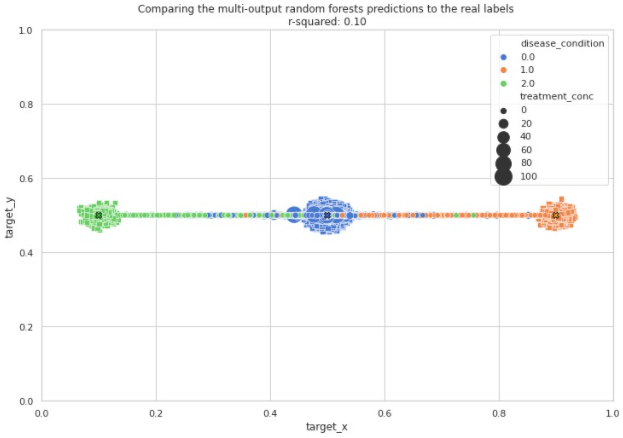


Figure 4: A challenging plane modelling. Active centroid is placed in the center, between the irradiated and the mock.



# Discussion

The results of our limited-scope study have managed to reproduce the conclusions of [[2](#r2)]. We were able to achieve that with limited domain knowledge, and no significant preprocess of the data. Moreover, we used Random Forest regression as “out of the box” algorithm, with no need for special adjustments. Equally important, our initial naïve modelling of the problem space was sufficient, establishing that the RP methodology can be based on simple and intuitive modelling. This is in contrast to the original RP usage which is based on expert-generated model.

# Future work and possible improvements

Future work should be made in several distinct directions. Firstly, the approach should be refined. Secondly, more thorough biological analysis should be done. Lastly, a comprehensive study of the approach and sensitivity analysis should be accomplished.

There is a need for a more decisive efficacy measure. A method of combining our proposed metrics may be used for this task. Another approach would be to utilize a method suggested by the original RP paper [[3](#r3)], proposing the usage of Kernel Density Estimation (KDE) on the RP created by the trained model’s predictions. This approach might also assist in modeling the drug dosage effect (dosage-based trajectory). A different method that may enable the prediction of treatments dosages trajectories is using 3D modeling. The modelling process remains as we described above, however, the (known) dosage is combined with the assigned target coordinates. A regression model is trained to predict both coordinates and dosage levels.

There are more interesting topics to explore using our method from a biological perspective. For instance, an examination of Vero and HRCE cells separately can be performed. Are these different cell types correlated with their response to a specific treatment? Are their trajectories the same? Does a specific treatment affect one cell type more than the other? Additional important biological questions that can be asked require domain expert’s analysis. For example, does the trained model’s predictions correlate with the known stages of active COVID-19? expert analysis is needed to evaluate the cell states in the predicted trajectories.

From the standpoint of data science and machine learning there are significant questions that need to be explored. Prominently, generalization of the proposed method to other tasks and datasets. In addition, the usage of different regression algorithms should be studied to examine the algorithm effect on the biological conclusion.

# References

1. World Health Organization. (n.d.). WHO coronavirus (COVID-19) Dashboard. World Health Organization. https://covid19.who.int/
2. Heiser, K., McLean, P. F., Davis, C. T., Fogelson, B., Gordon, H. B., Jacobson, P., ... & Gibson, C. C. (2020). Identification of potential treatments for COVID-19 through artificial intelligence-enabled phenomic analysis of human cells infected with SARS-CoV-2. *BioRxiv*.‏
3. Szkalisity, A., Piccinini, F., Beleon, A., Balassa, T., Varga, I. G., Migh, E., ... & Horvath, P. (2021). Regression plane concept for analysing continuous cellular processes with machine learning. Nature communications, 12(1), 1-9.‏