Group F Peer Review

General Comments

There have been decades of work in HIV research to simulate the virus well-enough to learn from it. Using cellular automata, researchers have simulated the spread of HIV within the body, but these computations are highly expensive. With data gathered from several of these leading simulations, authors sought to use common regression models to make predictions about the state of the simulation after 600 weeks, while only being given the first 300 weeks. This is a valuable insight as a well-trained model could be used to dramatically increase the speed of test results for HIV patients. The paper is generally clear and appropriately explains complex topics and tools used.

Specific Comments

Major Comments

1. The simulations outlined in the introduction need to include some detail about why they are reliable or appropriate for simulating the spread of HIV. I happen to know that some of these simulations are products of award-winning research in public health, but other readers might not. Even saying that the models are accurate and that the explanation will come later on would be sufficient for the introduction, but some note need to be made on why we can trust the simulations that you gather all your data from are reliable or accurate. If we do not have any sense for the quality of the simulations, then we have no sense for the meaning of your results in a real-world application.

2. Section 2.2 details what has collected from each timestep. While it is valuable to say what sci-kit identified as important features, it would also be wise to include a discussion about features we expect to be important given previous research. In the future, we should consider our accuracy with different feature sets in order to compare and contrast results.

3. Section 2.3 on regression outlines each regression technique used to make predications. Each has its own adequate overview, other than Lasso. The Lasso section could use considerably more depth as it is important that you choose between LassoLars, LassoCv, and LassoLarsCV (all from sci-kit learn) and explain why you chose that implementation.

Minor Comments

4. Your goals, or potential benefits from your research, are not clear until reading the conclusion. The conclusion adequately explains why having an accurate regression model would be useful to HIV treatment plans, but this needs to be presented to the reader as early as the introduction. This could really help to tie up the paper together.

5. Some awkward phrasing at the end of the first paragraph in the introduction. “Much time to generate results” sounds clunky and unnatural. Perhaps just mention that the simulations are highly computationally expensive

6. The start of the second paragraph in the introduction could use some rephrasing. You don’t bring machine learning techniques to make predications. Maybe a better way to phrase it would be building regression models to make predications.

7. Again, some awkward wording in the results section. “How different regressors were at…” could benefit from some rephrasing. Something like, we compare the accuracy of different regressors when predicting the final state of each model, would be more appropriate here.

8. Python is mentioned in the results section for some reason which seems very out of place. If you think it is necessary to say python was used, maybe it would be better to mention earlier. Although, I do not think it is necessary in the first place.

9. To go along with emphasizing the goal of your research, I think it is also necessary to emphasize the importance of the research that preceded your own. As early as the introduction you should explain why we care about finding a less computationally intensive way of evaluating HIV simulations. Small edits like this throughout the paper should make for a very clear read.