Response To Reviewers Comments

Dan MacLean

21/06/2021

Dear Editor,

Thank you for taking the time to send our manuscript for review. Please find attached a list of revisions in response to the reviewers comments that you provided.

We hope that you will find these satisfactory

Sincerely

Prof. Dan MacLean

## Reviewer 1

1. 'The language in the manuscript needs close attention. For example, the sentence "Deep learning approaches require positive and negative examples from which to learn - here a list of sequences annotated as an effector or not.'

The manuscript was written by a native English speaker and we believe in general that the text is clear, as the same reviewer points out the authors explained the concepts very clearly. We have modified the sentence mentioned to read Deep learning approaches require positive and negative examples from which to learn. We used a list of sequences annotated as an effector or not

2. 'This sentence "The larger and more accurate the list the more sensitivity a model can obtain." requires a citation of other works that explore the change in sensitivity with increasing training examples.

No such citation can be made as no study exists. This is because statement is one of a trivial fact about training set size. It is considered self-evident in the field that the more training examples you have, the better predictions can become. We have rephrased to It is generally held and expected that he larger and more accurate the list the more sensitivity a model can obtain.

3. 'It will be nice to conduct a Wilcoxin's [sic] test to ensure that the difference in performance of the different classifiers are significant.'

We understand the impulse behind the suggestion but believe that this is not appropriate for the data presented. The per data set classifications can only be tested once, as they result in the same output for the same input every time, there is no experimental variability. Furthermore, it does not seem really informative to run hypothesis tests on these data, if the reviewer is suggesting pooling the accuracies we’d be testing the hypotheses that accuracy over all test data sets we used the tools on is equal. If one was significant over all data sets it would seem to suggest that the tool was ‘better’ overall. Our tool would likely come out better over all the datasets, but such a result would mask the more important results about relative performance across kingdoms and different data sets that could guide a user better. The confusion matrices are the best indicators of relative performance characteristics for a given test data set and we use those. Using a single hypothesis test across all data would leave us open to the justified criticism that we had cherry picked a single summary. This is not our intention, so we suggest not to perform this summary statistic and rely on the more explicit, detailed statistics that we have.

4. 'It will be good to mention the location of the R package next to the sentence "We have produced an R package that will allow other scientists to easily classify their own sequences of interest."'

We have added the link to the R package homepage at the point the reviewer suggests, in addition to its existing places in the manuscript.

5. 'Can this software be used to predict if a protein is an effector or not? You mentioned how the data was curated for each of the kingdoms. I wonder how the negative data would be constructed if you were to set it up to simply predict if a protein was an effector or not.'

For the first part, yes, this software can be used to predict if a protein is an effector. Our entire manuscript, including the title, and this reviewer point this out in their opening statement In this manuscript the authors have explored a novel technique to predict effectors We believe that this point is covered well enough if the reviewer did in fact notice it.

For the second part, we are not sure how to proceed. This seems to be a thought from the reviewer that we have already covered. The reviewer seems to be wondering how we might set up negative data in the case that we want to predict if a protein was an effector. As we do want to predict whether proteins are effectors, and did need negative data, we set it up as fully described in materials and methods, which the reviewer has noted we did. We do not think anything further needs doing for this point.

## Reviewer 2

1. 'It is a common and most essential practice to train and test methods using different datasets. Authors used 60%, 20% and 20% for train,test and validation. Is the selection random? How many times authors repeated it with different combination of train,test and validation tests.'

The selection was random at each model, we have updated the section Encoding and Subsetting Sequences to reflect this now reading Encoded sequences were split into random taxon specific training, test and validation. The process over hyperparameter scans was repeated 3000 times, resulting in 3000 models being tested. This is described in section Model Selection and Training

2. 'Did the authors checked the performance of the proposed method against the datasets used in the existing methods?'

Yes, we did. This is described in the dedicated section Comparison With Other Classification Software and highlighted as We also evaluated the deep learning models we have developed on the training and hold-out validation sequences used to train the previous methods. in the second paragraph. Figure 10 shows the results of this.

3. 'Including a table that shows the sensitivity, specificity and accuracy of the proposed method and existing methods on test/validation dataset will be needed.'

The data are already described in confusion matrices in Figures 6, 9 and 10 and in Figure 8 along with the F1 score.

## Reviewer 3

1. 'Did authors have performed 3 fold or 5 fold cross-validation?'

We did perform five fold cross-validation in the model training, we have added Five-fold cross validation was used in all instances. to the methods section.

2. 'Authors have selected multiple hyperparameters for each classifiers, and mentioned that they have further chosen some of the hyperparameters to take multiple values and fixed rest of the hyperparameters for each classifier. What criteria were in mind or how authors have decided that which hyperparameters should be fixed and which not? This should be elaborated in methods section. GRID search - explain that.'

As described in the methods section and section Model Selection and Training (paragraph 2 and 3), we selected hyperparameters likely to have a strong effect and for those selected random values within likely useful ranges, running a model for each combination, a so-called random search. We present the range and values that we used in the associated supplemental tables. We are not sure why the reviewer has suggested we need to explain a grid search as it is not something we performed. We performed a random search of the space which we have described.

3. 'Line 35 on page 4,missing figure number.'

This is corrected.

4. 'Supplementary tables are not in correct order, please correct them.'

This is corrected.

5. 'In figure 11, it is not clear what does left, middle and right panel for each taxon represent. Make it more descriptive in figure caption'

We have updated the caption to read: Grad-CAM activation profiles of CNN-LSTM models when run with all datasets showing the region of the sequence the model is most able to use to classify (leftmost panel). The most active part of the sequences are shown in logo plots of sequences extracted from the activation region within the dotted lines for effector data (centre panel) and non-effector data (rightmost panel). A: Models trained on bacterial sequences; B: Models trained on fungal sequences; C: Models trained on oomycete sequences.