Dear Dr. Radivojac and Dr. Rost,

General comment: reviewer comments should be in italics and your response in plain text.

Thank you for the very helpful comments on our submission. We have responded to the specific reviewer comments below, but in addition would also like to comment on the work of Elkan and Noto that you directed us to.

Elkan and Noto provide a framework for learning without negative examples, which relies on one key assumption: The probability of a positive example being labeled is independent of the example in question (labels are selected at random from the positive class). Unfortunately, in the protein function prediction context, this assumption does not hold, as new annotations are often added from homology links, and thus the samples in question are far from independent. In addition, their method requires the estimation of a constant c, which is the probability of an example being labeled given that it is positive. They determine this through a classifier which attempts to predict if an example will be labeled or not. In effect, we have computed just such a classifier in earlier work (Reference #4 in our submission), specifically the classifier which treats all non-positive examples as negative, is precisely the same as a classifier that tries to determine if an example will be labeled or not. We showed in this work that such a classifier achieved worse results that one utilizing our negative examples.

While it is entirely possible that the work of Elkan and Noto could be successfully adapted to function prediction, despite the *a priori* violation of the underlying assumption, we believe that in light of this and our other points above, as well as the demonstrated contribution of our negative examples to function prediction accuracy, which we have introduced in this revision, that our work makes a contribution above and beyond the results of Elkan and Noto. We leave it to those authors to further develop their methods and apply them to the protein function prediction problem.

The individual responses to reviewer comments can be found below, followed by a detailed changelist of modification we have made to the paper.

Sincerely,

Noah Youngs

**Reviewer #1**

“To demonstrate the effectiveness of the negative protein function examples, the authors should train a machine learning predictor on the negative examples curated in this work and the randomly selected negative examples separately. Then the authors can compare the performance of the two predictors and discuss the effectiveness of using the negative examples curated in this work. “

-- We have included a new figure (Fig 4), along with a section in the main text (section “Case Study: Improving function prediction in Human, Mouse, and Yeast”). That utilizes the most successful negative example methods to perform function prediction. The results follow closely the results of our negative example algorithm validation, and demonstrate that our negative example selection methods cause a significant boost in function prediction accuracy over random negative selection.

**Reviewer #2:**

“PloS-CB is a journal which is read by a broad range of computational biologists. I'm afraid that the methodology described in the first paragraphs of the Results section is too succinct for someone not directly in the field to understand. It would be helpful to have a flowchart, or a more elaborate explanation of the method used.”

-- We have expanded upon the results, section:”Evaluation of Negative Example Quality”, in order to make our methods more broadly understandable.

“In Figure 1 it appears that, overall, even with random predictions, the average (I assume "average" means "mean") number of false negatives per negative predictions is very low for practical purposes. At worst, in the Biological Process category of the human genome, the Random average is 7 false negative predictions per 3000. This drops to about 1.5 / 3000 with the best performing method (SNOB). I would appreciate if the authors could comment on the significance of the differences between the various methods given such a low false-positive rate in the first place. Perhaps emphasizing the differences between prediction on terms of different specificities would help clarify the information conveyed in Figure 1 (see also comment  below).”

-- We have included additional validation by way of utilizing the different negative example methods to perform function prediction (figure 4 and section “Case Study: Improving function prediction in Human, Mouse, and Yeast”). We believe that the large performance difference demonstrates the significance of the reduction in false error rate. We have also added additional text in section “Evaluation of Negative Example Quality**”**, emphasizing that the error rate we measure is the observed error rate, rather than the true error rate.

Because we can observe errors only on newly annotated genes, the upper bound of the error rate is the average number of new annotations per category. For the example you mentioned, the upper bound on the Biological Process category would be an error rate of 45.4 when choosing 3000 negative examples, meaning that the random baseline achieved16% of the maximum observable error, while SNOB achieved just 3%. The main reason for the low cardinality of the error rates in BP, is that there are many categories which only received one new annotation, and thus have a maximum error rate of 1.

“On term specificity: the authors define term specificity by the number of proteins in GO annotated by that term. Would that be for a leaf term? Or for any propagated term? So, for example, of a protein is labeled as "alkaline phosphatase" it also has "protein binding" implied (propagated) term. This leads to a broader question: Is each protein annotated with a GO subgraph, or only with a set of terms? The answer to this question is important, since it determines whether specificity is determined by frequency of leaf terms, or by actual correlation with the GO term hierarchy in the GO graph. “

-- The number of proteins is counted by propagating all terms, as defined by the rules governing the GO hierarchy (GO obeys the “true path rule”, meaning that an annotation to a particular term implies annotation to all parent terms). Thus each protein is annotated with its entire GO subgraph. We have added text to section: “Data Processing” in Methods, to clarify this point.

“Also, in Figure S1, the false negative rate is high for the prediction  terms of low specificity, but low for terms of high specificity. This is expected, as low-specificity terms. The differences are quite dramatic, which illustrates that the averaged results, as shown in Figure 1, are not very informative. The authors should address that.”

-- We have added text to address your point (section “Performance of Negative Example Methods in *Homo sapiens”*).

“The number of false negatives for the prediction of GO term "RNA binding"  Figure 2) seems significantly higher (than the ones in Figure 1. This warrants explanation which I could not find in the text. Also, the authors made the Y-axis in Figure 2 too large to be able to be read meaningfully. Each tick is 10 False Negatives, and all the methods, except for 1-DNF have a false negative prediction rate of under 10. I suggest the authors rescale that figure. Also, any idea why 1-DNF performs so poorly in that category? “

-- As described in section “Performance of Negative Example Methods in *Homo sapiens”*, the RNA binding term evaluation includes many validation proteins from the Batlz et al paper, which allows for a much higher maximum error rate than other categories for which we have fewer new annotations. We have rescaled the figure as per your recommendation.

“Figure 3: I could not see the 1-DNF prediction well (it's barely visible near the origin). “

-- We have changed the marker of the 1-DNF prediction to make it more visible

“Figure 4: panels are not labeled A,B & C. I assume it is left-to-right.”

-- We have made the labeling explicit (the figure is now figure 5).

**Reviewer #3:**

“1) The first thing is easy to address. Authors should convey better the message that selecting negative examples is not exactly the other side of the coin of the prediction of positive examples. In other words, it is not exactly providing “negative examples” by taking all of the GO terms minus positive annotations. “

-- We have added text to the introduction to further emphasize this point.

“2) Selection of negative examples are given as a list of ranked proteins for a particular GO but I ask to:

a. provide a score (raw score or better a p-value or something like that)

b. provide “negative” GO for a protein and not only the GO with its list of “negative” proteins. This would be really helpful.”

-- With regard to a), we will be incorporating a score for each negative example in the second iteration of our NOGO database. In terms of b, this amounts to choosing a cutoff for each GO category, and then organizing the resulting negative examples by protein instead of by function. We believe this is best left to the end-user of our negative examples, as they can utilize the validation plots we provide for each function to select a cutoff that fits their particular accuracy needs.

“3) Authors show that the performance of SNOB (H. sapiens) is effective and performs better than other methods in predicting negative examples for more general GO categories i.e. GO terms with a low information content as they are close to the root of the GO graph and are more frequent. The performance decreases when GO terms are rare and subsequently more specific. What I suspect, on the contrary, is that methods have a general tendency to perform better, in any case, when GO terms are rare or have a low frequency if compared with those that are highly frequent. As an example one can check GO:0005515 “protein binding” and GO:0019901 “kinase binding” which are generic/highly frequent and more specific/less frequent respectively.

Authors must explain what are the real effects they observe in the results when considering the strong biases in the frequencies of some GO terms which can vary a lot from very low to very high. It is not surprising that the number of erroneous negative examples increases proportionally with the higher frequency of the selected GO. Similarly, it is not surprising to perform well when the GO term is rare. This brings me back to what I said earlier, namely the need to have a score and not just an ordered list of proteins for a selected GO. This can help the user to choose on his own the desired threshold depending on the GO term (generic or specific) the user is interested in. “

-- We have added additional text to section “Evaluation of Negative Example Quality” to further clarify the effects of GO term frequency on performance. We believe that our methodology of comparing to a random-selection baseline does a good job of mitigating frequency-based evaluation biases, as the relative performance of a given method to the random-baseline is agnostic to the frequency of the term in question.

“4) Finding a benchmark is extremely difficult and authors have had to deal with the same problem everyone has to face when assessing one’s method. Nonetheless, I find a weak evidence reporting results for single GO terms and spend an entire paragraph. I suggest limiting this part and focus on general trends and what I ask in point 3) “

-- As mentioned in the text, we believe that the two specific GO terms we have focused on are worthwhile, as they represent the most complete sets of annotations that currently available for any GO terms, and thus the closest to a fully-observed validation set.

“ “ ….. and (iii) using genes with annotations in sibling categories of the category …” (iii) what does this stand for ? “

-- This was a typo and has been removed.

“Some typos like “Rochhio” in First page of Introduction”

-- We have fixed all the typos we could find.

**Reviewer #4:**

“1. They authors include annotations with the IEA evidence code in their training and evaluation sets.  These are predictions that are not experimentally verified, although they have been shown to have a high accuracy.  So including them is not unreasonable, but doing so may also bias the results. To examine this potential bias, the authors would need to repeat the same analyses they have already performed, but excluding IEA annotations in the training and test sets.  The two main sources of IEA annotations are published (Burge et al, Database: bar068, 2012) or documented (<http://www.uniprot.org/program/automatic_annotation>). For selection of negative examples, at least one relevant point is that these methods apply the same set of GO terms to a large number of different proteins.  Thus, for selection algorithms that include co-occurrence of GO terms or treat each protein as a single "document", the strongest signals are expected to derive from IEAs.  The authors may find other ways to deal with the resulting spurious correlations in both the training and test data, such as treating all proteins annotated by the same InterPro model as one observation rather than N independent ones.  But at the very least, they should repeat their analyses but excluding the IEAs.”

-- We have repeated our same analysis, but excluding the IEA annotations from the validation process, in order to remove any potential bias they might introduce. The results mirror those of our original analysis, and are included as supplemental figures as well (Figure S4, Figure S5).

“2. It would be very helpful to give some examples of negatives that were selected by different algorithms but assessed as being incorrect in the test set.  An analysis of the kinds of errors that are made by different algorithms can illustrate the differences between the algorithms, and the caveats that the scientific community should be aware of when using the predicted negative examples from them.  Are there cases where the predicted negative examples find errors in the experimentally-supported GO annotations?”

--. Because our methodology treats existing annotations as the ground-truth (at least as far as it is observable), we do not predict any negatives which currently have an experimentally-supported GO annotation in the category in question. Analysis of specific proteins that were misclassified will be a subject for future work.

“3. Following from the previous two points, the authors should use the negative examples derived from the IEA-excluded training set, and find specific IEA (predicted) annotations that were inferred to be negative examples.  This would be a test of using the predicted negative examples to assess the quality of predicted positive examples.  How many are there? Is there reason to believe some of the predicted positives might be incorrect?  Again, specific examples would be helpful. One could even argue that algorithms that suggest a greater number of negative examples that are actually predicted by IEA methods are more relevant, as these examples would help classifiers to be more discriminative.”

-- We agree that utilizing negative example prediction to assess IEA prediction quality is a potentially interesting avenue of research, however the present work seeks to demonstrate the accuracy of several negative-prediction methods. We hope that these methods will be employed in a variety of applications in the future, including the one that you have just mentioned.

“4. As the positive examples in the GO database (the basis for the negative examples described in the paper) are always evolving, it would be more useful for the community to access to the code for generating negative examples, than a database.  The authors should provide the code and usage documentation for their algorithms, and links to third party code when appropriate.”

-- We have posted the code to our database website, as well as documentation and input format instructions, to allow others to generate negative examples from future GO releases.

**Changelist**

* Added Figure 4
* Added section: “Case Study: Improving function prediction in Human, Mouse, and Yeast” describing Figure 4
* Added section “Function Prediction Implementation”to methods section.
* Added text to the discussion commenting on Figure 4
* Added text to the introduction describing the case study we did for figure 4
* Added text to the results section to further explain methods
* Added text to section “Evaluation of Negative Example Quality” to further clarify the bounds on the size of the observed error rate.
* Added text to section “Data Processing” clarifying how GO terms are propagated
* Added text to section “Performance of Negative Example Methods in *Homo sapiens”* to address the impact of the specificity of terms on the evaluation metrics.
* Rescaled figure 2
* Changed 1-DNF marker in figure 3
* Labeled panels in figure 4 (now figure 5)
* Added supplemental figures S4 and S5 depicting performance without IEA results in the validation
* Added text to section “Evaluation of Negative Example Quality” explaining the impetus for supp figures S4 and S5
* Fixed typo in introduction (the iii)
* Fixed misspelling of Rocchio in introduction
* Added text to the introduction to clarify difference between predicting negative and positive.
* Added text to section “Evaluation of Negative Example Quality” further addressing performance and frequency correlation