Supplementary Materials: The impact of incomplete knowledge on the evaluation of protein function prediction

Yuxiang Jiang, Wyatt T. Clark, Iddo Friedberg, and Predrag Radivojac

June 14, 2014

1 Incompleteness of protein function annotation

In Supplementary Figure 1 we provide histograms of the number of new annotations in Swiss-Prot by comparing Swiss-Prot experimental annotations from January 2011 to those from January 2014. We also refer to the number of these newly added terms as the *size* of new annotations.

2 The scenario of consistent performance on new annotations

We have shown in the main manuscript that the new annotations can lead to either positive or negative changes of the F-measure depending on several factors: (*i*) the precision and recall of the predictor that is under evaluation; (*ii*) the performance of that predictor on the new annotations (rc_δ); (*iii*) the size of incomplete annotations ($\rho = \delta/\gamma$, where $\delta = |T' - T|$ and $\gamma = |T| = tp + fn$).

Here we investigate the impact on F-measure under the assumption that the recall of a predictor on the new annotations is the same as recall on the complete-data annotation, i.e. $rc_\delta = rc'$, which in turn leads to $rc_\delta = rc$. The changes of F-measure can now be expressed as a function of the original performance (pr, rc) and the level of incompleteness.

$$\Delta_{F_1}(pr, rc, \rho) = F_1 \cdot \rho \cdot \frac{2 \cdot rc - F_1}{2 \cdot rc + F_1 \cdot \rho} \quad \left(pr \leq \frac{1}{1+\rho}\right),$$

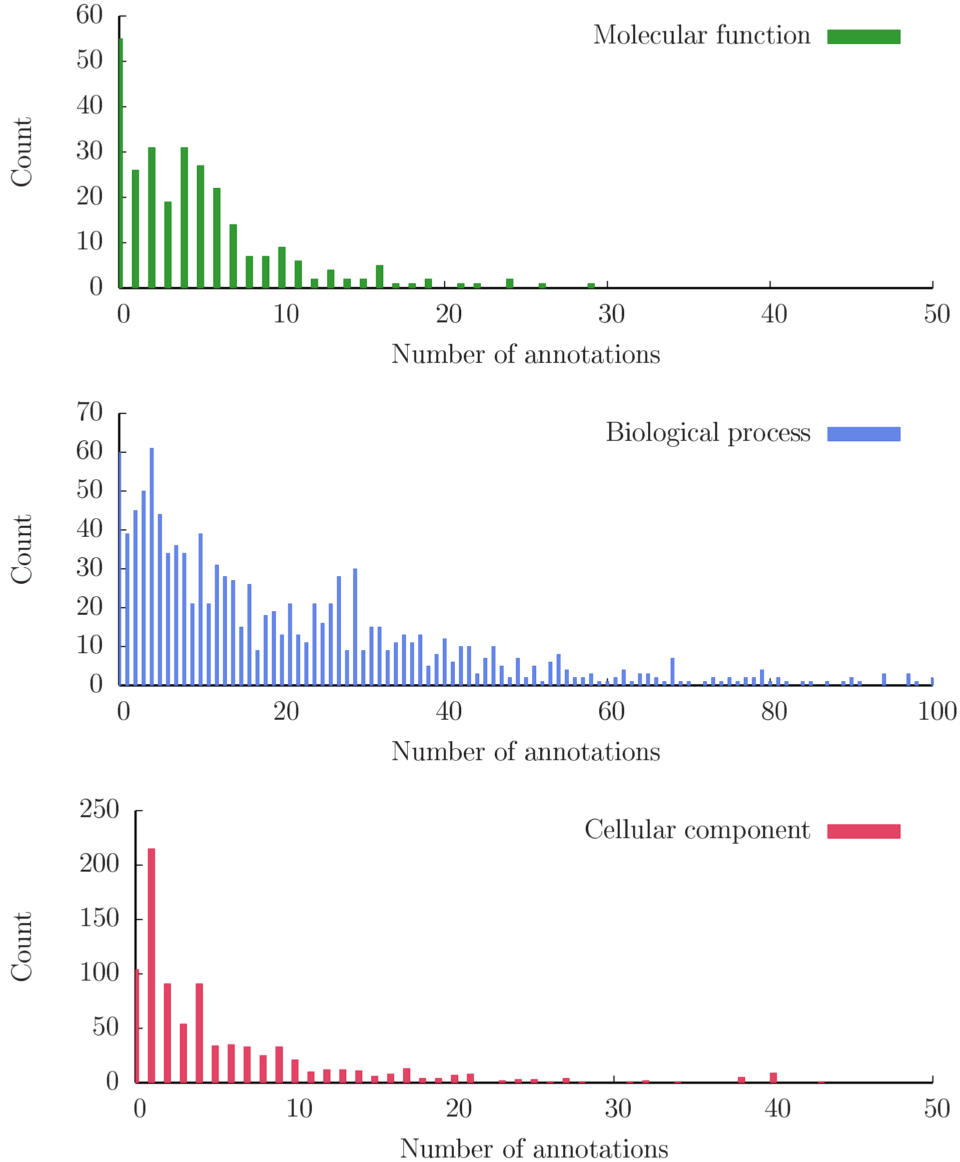
where F_1, as before, is the original F-measure. Note that $pr \leq (1+\rho)^{-1}$ directly follows from the condition that $\beta \leq fp$; that is, the number of false positives in the complete-data confusion matrix must be non-negative. Here we visualize (Supplementary Figure 2) the impact by fixing the incompleteness to some specific levels relative to the original annotation: (*i*) half the size of the original annotations; i.e. $|T' - T| = \frac{1}{2}|T|$; (*i*) the same size as the original annotations; i.e. $|T' - T| = |T|$.

In the first simulation (Supplementary Figure 3), we fixed the number of original and new annotations to 50 but added some randomness to rc_δ by letting the number of correctly predicted terms on the new annotations follow a binomial distribution: $\beta \sim \mathrm{Bin}(\delta, rc)$. Blank regions in Supplementary Figure 3 are infeasible regions, i.e. they indicate that none of 10000 trails resulted in a valid β.

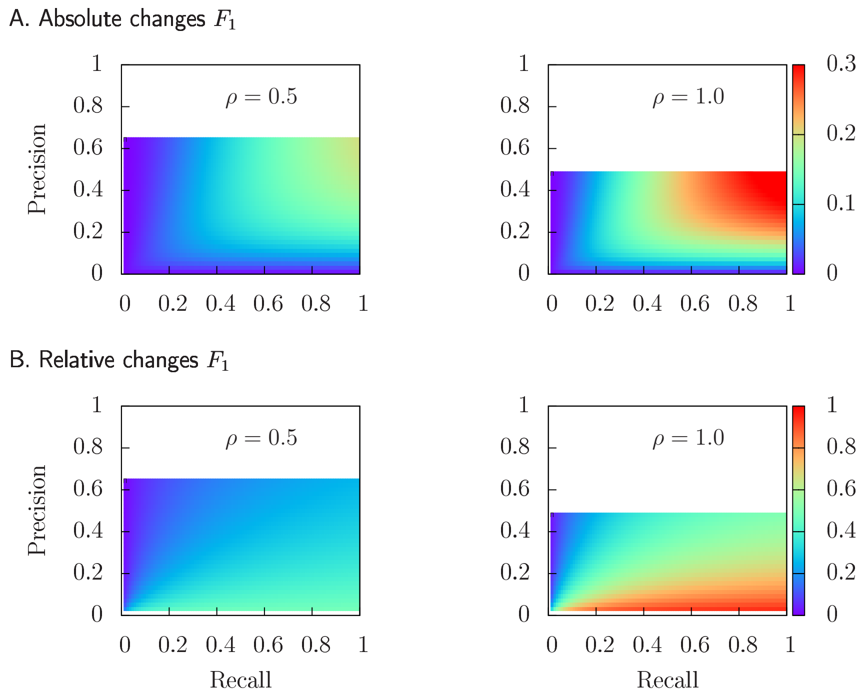
3 The scenario of decreased performance on new annotations

In previous simulations (including those in the main manuscript), we assumed that the performance of a predictor on the new annotations remains the same. However, it is relatively difficult for a predictor to maintain its performance on the new annotations because, in many cases, the predictor was trained on a data set with limited or even biased knowledge and new annotations. Therefore, it is also interesting to simulate the impact under the scenario that the recall decreases on new annotations.

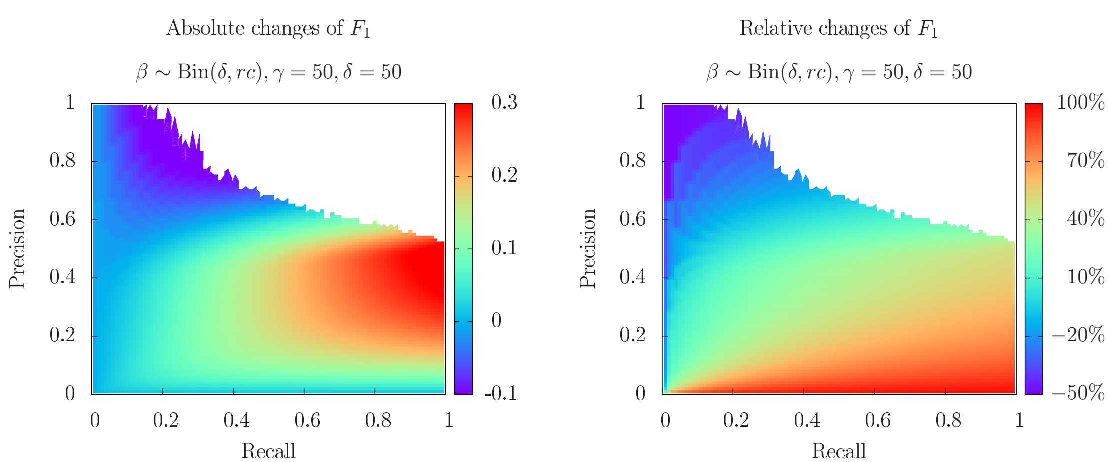
Supplementary Figure 4 shows the simulation result under the same setting as described for Supplementary Figure 3 except that we let $\beta \sim \mathrm{Bin}(\delta, \frac{1}{2}rc)$. By comparing these two experiments, we note that the impact on F-measure under an arguably more realistic setting is much smaller.



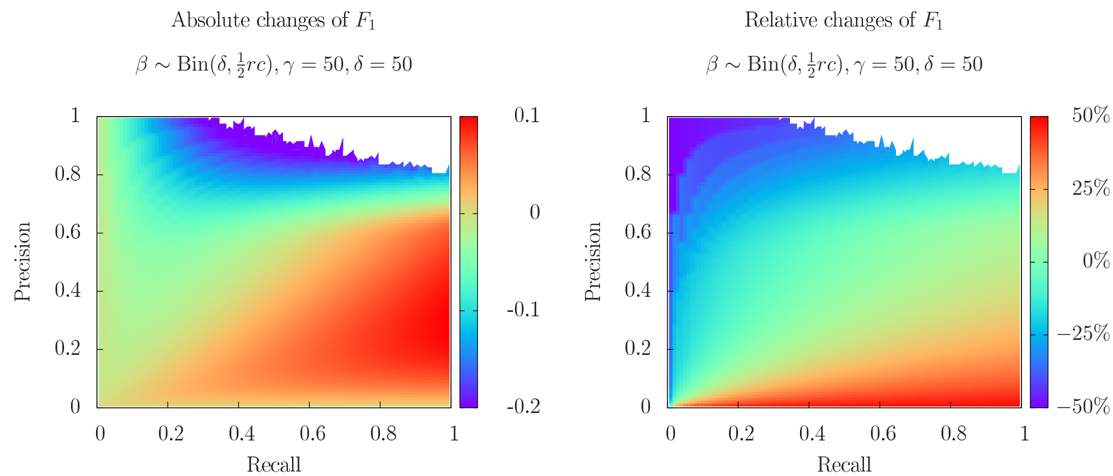
Supplementary Figure 1: Number of new annotations (new GO terms) after a three-year period from January 2011 to January 2014 for three GO ontologies. Note that 21 proteins with more than 100 newly added terms from Biological process ontology are excluded in the histogram for visualization purposes. The numbers include entire subgraphs; thus, one added GO term can result in a size of new annotations to be significantly larger than one.



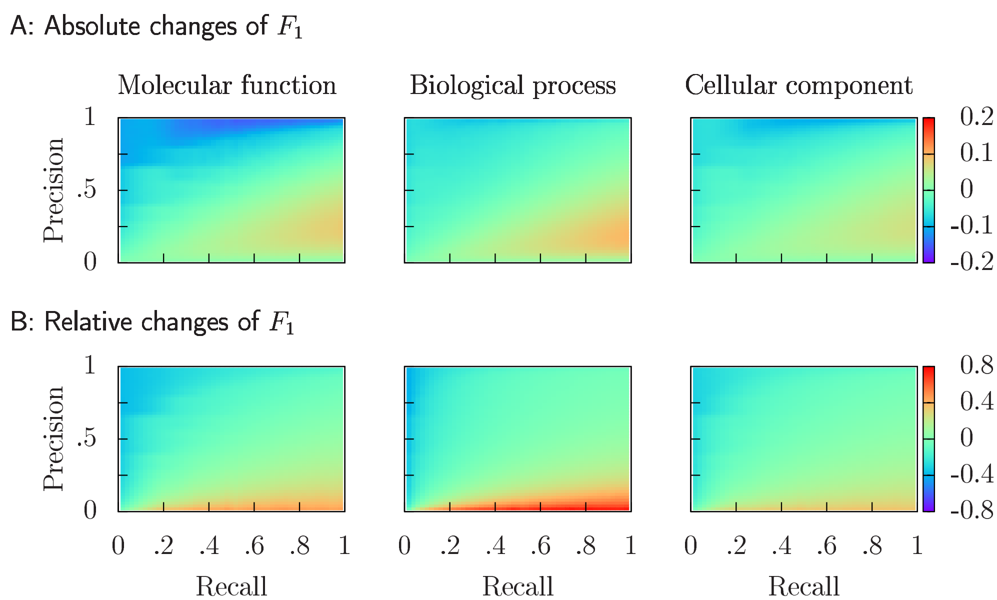
Supplementary Figure 2: Impact of new annotations on F-measure under the assumption of $rc_\delta = rc$ with different size of new annotations. $\rho = \delta/\gamma$, where $\delta = |T' - T|$ and $\gamma = |T|$. (A) Absolute changes; (B) Reltaive changes.



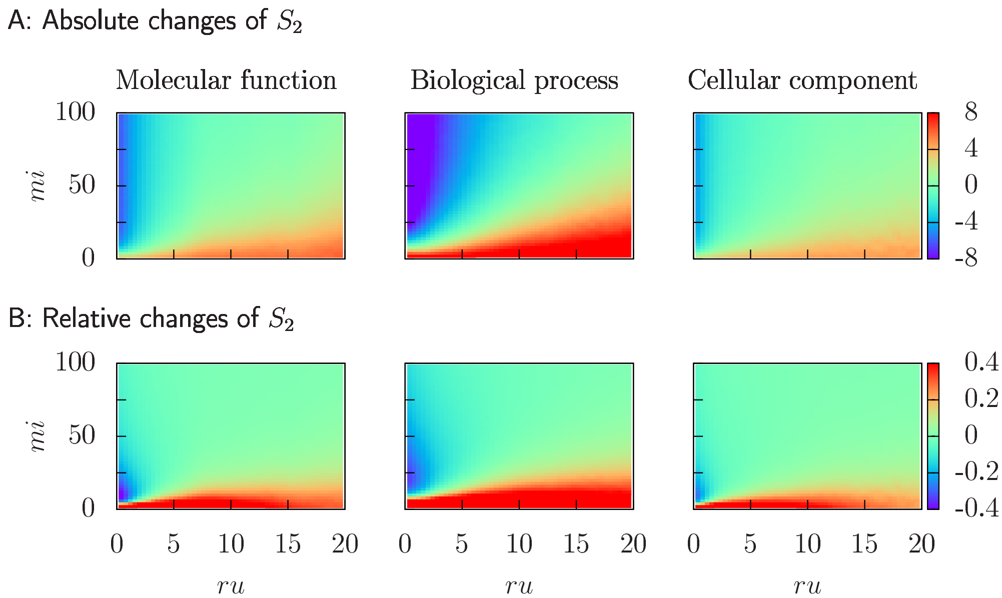
Supplementary Figure 3: Impact on F-measure by assuming the performance on new annotations remains the same as that on the original annotation. This simulation corresponds to $\rho = 1.0$ in Supplementary Figure 2.



Supplementary Figure 4: Impact on F-measure by assuming the recall on new annotations decreases to a half of that on the original annotation.



Supplementary Figure 5: Simulated changes of F_1 with decreases recall on new annotations. A: absolute changes, and B: relative changes, as a function of precision and recall estimated on incomplete data.



Supplementary Figure 6: Simulated changes of S_2 with decreased recall on new annotations. A: absolute changes, and B: relative changes, as a function of misinformation and remaining uncertainty estimated on incomplete data.

Supplementary Figures 5-6 show the simulation results of the impact on F-measure and semantic distance, respectively. Note that they are under the same setting as simulations in Figures 3-4 from the main paper except that we let $\beta \sim \text{Bin}(\delta, \frac{1}{2}rc)$ and $\beta/\delta \sim \text{B}(\frac{1}{2}(\gamma - ru), ru)$ in order to simulate a decreased performance on new annotations. It is interesting to see that the impact on F-measure under this changed setting becomes minimal while the impact on semantic distance does not differ as much.