# Response to reviewers

## Response to Reviewer 1

This manuscript describes a computational design and experimental testing of a library of glycoside hydrolase mutants, constructed to understand the sequence-function relationships in this enzyme. While this work represents an interesting contribution to the field of protein modification, the authors’ claim in the Conclusions that “[t]his data set will be invaluable for the development of computational enzyme engineering algorithms” seems to be overly optimistic.

One of the problems is that the data and the manuscript have not been carefully prepared and proofread. There is a number of discrepancies between Fig.2 and Suppl. Table 1, which are supposedly describing the same set of mutants and contain the key results of this paper. These discrepancies include: mutants Q284R and W399C appear in the Table 1, but not in the Figure 2. C167A appears twice in the Fig.2. Position 17 and 407 have two different wild type amino acids assigned to each in both the Table and the Figure (E and S for 17, and W and N for 407). These errors can be easily spotted, but they raise a question about quality of the rest of the data, which cannot be readily checked by a reviewer. In the References section refs. 2-6, 21, and 25 are not formatted correctly with the key information missing.

**We are grateful to Reviewer 1 for finding these errors in our manuscript. In addition to the discrepancies identified by the reviewer we have cross-checked all of the data against the oligo sequences, raw data, preliminary data sets, sequence alignments, and structures. We are confident in our data set as now presented. All of the errors derived from manual labeling, manual generation of the heatmap, and/or numbering discrepancies between crystal structure and gene sequence numbering. A few redundancies were found in this process (due to crystal sequence numbering issues) and consolidated. We have also corrected the references, including the missing author information.**

In the last sentence of the Section “Observed sequence–structure–function relationships in BglB” a preposition is seemingly missing. The authors should carefully recheck the entire manuscript.

**We have revised this sentence (and finely checked the rest of the manuscript) to remove the error.**

## Reviewer 2

1) The reliability of Rosetta structural predictions. The authors’ premise is that molecular modeling using Rosetta has not been a reliable predictor of function as currently applied. This seems supported by the data presented here. Many of the 104 substitutions made in BglB were chosen based on Rosetta predictions of neutral or positive impact on presumed transitions state binding. Analysis of the results shown in Suppl. Table 1, however, shows that most substitutions have a negative impact on activity. Indeed it is difficult to conclude that the impacts of these selected mutations are significantly different from the Alanine substitutions to all active-site proximal positions. Perhaps the authors could speculate on the reliability of the Rosetta method to predict functional relevance, particularly since they speculate on the mechanistic implications of function-correlated Rosetta model features.

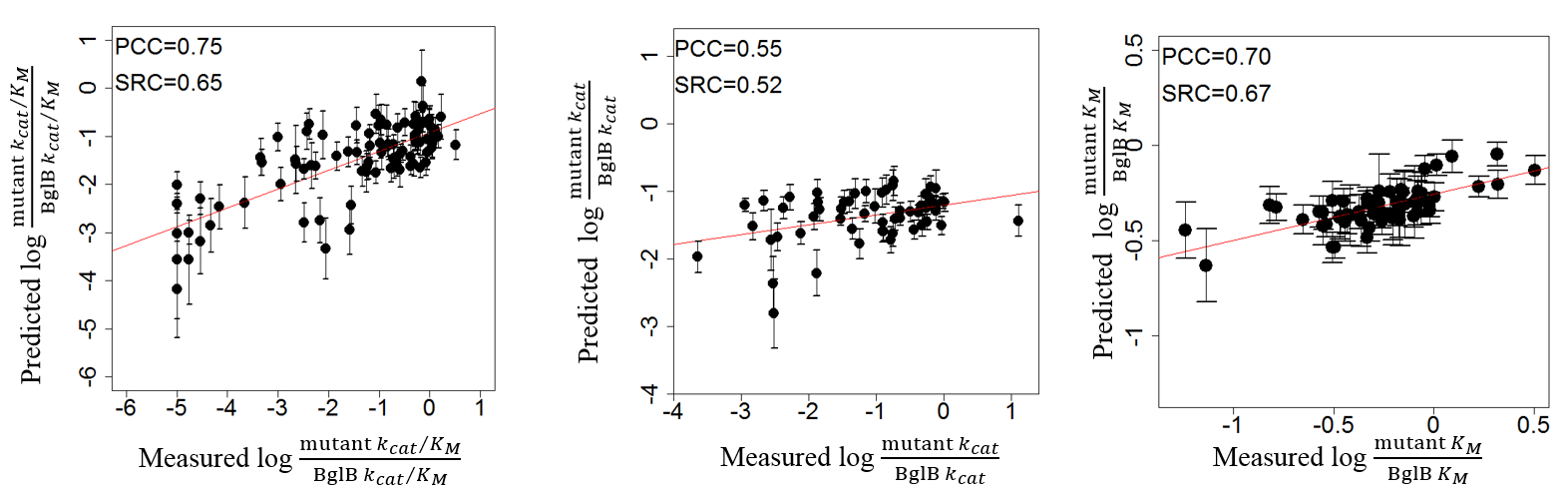
2) Lack of prediction of R240A. The one substitution that yields a significant improvement in activity, R240A (presumably from the Alanine scan), is not predicted to be positive based on the regression model shown in Figure 4. The overall method does not seem to yield significant information on how to improve activity, with nearly all variance being to the negative. If the authors suggest this as a method to engineer new activities, then some example of directed improvement should be shown. This could be combining current mutations based on the model to create further improvements or adding new substitutions predicted by Rosetta to give such improvements.

**We thank Reviewer 2 for helpful comments about the conclusions drawn in our manuscript. In response to point 1 and 2 together: The variance in activities being almost entirely negative (worse than wild type) this has more to do with the data used to train the machine learning algorithms. From a structure-functional point of view, it is much easier to decrease the activity of BglB, even when using Rosetta to predict mutations that are compatible with the modeled transition state, and this bias is reflected in the data and in the predictive algorithm generated. In future efforts as we generate more mutants which will hopefully have increased activity we will begin to evaluate our predictive ability for the enhancement of function.**

3) Cross-validation of regression models. The authors use bootstrapping to create more reliable models from averages of many subsets, but it is not shown how well left-out subsets are predicted in cross-validation. This is somewhat captured in the error in prediction from subsets, it is important to show predictability of left out samples. Otherwise, it is not clear the degree to which the models are dependent on a small number of outlying samples. Correlation coefficients for prediction of left out data would suffice.

**We agree with the reviewer that evaluating the predicting performance on left out samples is crucial and actually this is what we have done in this work, although it wasn’t clearly stated in the manuscript. To clarify, we performed 10-fold cross-validation (CV) and evaluated the predicted performance on the left-out samples (generalization error) at each of the 10 runs. Then we repeated this procedure (i.e. the 10-fold CV) 1,000 times to randomize the sample distribution among the folds. That way, we reduce the effect of any bias for evaluating left-out prediction performance. The standard deviation over the 1000 10-fold CV trials for each datapoint is shown as error bars in Figure 4.**

**For completeness, we now also performed bootstrapping. We consider bootstrapping sets of size 2n, where n is the number of samples in the whole dataset (90, 80 and 80 samples for *k*cat/KM, *k*cat, and KM, respectively). Please note that this setting achieves an average coverage of 86.7% of the original data set in any given bootstrapping sample. The left-out samples were then predicted by an elastic net model training on the bootstrapping set. We repeated this procedure 1,000 times and then we averaged the prediction performance of the left-out samples over all runs. As shown in Fig. S9, the bootstrapping performance is similar to that of 10-fold CV that is depicted to Fig. 4 (slight variations due to smaller training/testing ratio).**

****

4) Error in reporting values, pg. 12. The reported PCC values for kcat and kcat/Km correlation to individual substitutions are switched in the text relative to the values in Supplemental Table 2.

**The order of mention of the kinetic constants has been switched in the revised manuscript.**