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.”

**Response:** 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) for 1,000 times to randomize the sample distribution among the folds. That way, we reduce additionally (to just 10-fold CV) the effect of any bias for evaluating left-out prediction performance.

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).

