Description of PON-Tstab method:

Benchmark datasets should sample the event space so that they could be used for method training and testing. Such datasets should fulfil several requirements. Relevance, i.e. that the cases represent the investigated feature, is one of them. A large number of methods have been developed to predict effects of amino acid substitutions on protein stability. Many of them utilize machine learning algorithms and have been trained with data from ProTherm database. When we examined cases in the database, we noticed a number of issues with the database contents and quality. Several issues emerged including some errors, but also features that had not been clearly communicated and which had caused data items to be used in wrong way by method developers.

We firstly checked thoroughly the details for variants in ProTherm database and corrected numerous problems. In the end, we had less than 50% of the original variants left. Out of these, 77% came from ProTherm, the rest are either corrected or new variants.

The final dataset contains 1564 entries from 99 proteins. This dataset is available in VariBench at <http://structure.bmc.lu.se/VariBench/stability.php>

With the new high-quality dataset we trained a novel machine learning predictor, PON-tstab, for amino acid substitution effects on stability and established a new baseline for variant stability prediction method performance. Our study revealed the importance of knowing and checking data that and their relevance when used for predictor development since predictors cannot be better than the data used to train them. Therefore, one has to be careful when using datasets collected by others unless they are properly documented and systematically compiled.

In total, 1106 features were collected to train a Random Forests based method for stability prediction. To eliminate redundant and non-relevant features, we used a combined greedy feature-selection algorithm with two steps: backward elimination and forward selection. Hence we got only on 8 features, including temperature, 1 conservation feature, 3 amino acid features, 2 neighbor features and 1 protein feature.

PON-tstab classifies the effect of variations on protein stability into three states: increase, decrease and no-change, also supplies a possibility\_score (from 0 to 1) for the classification result.

The web service is available at <http://structure.bmc.lu.se/PON-Tstab/>. And the manuscript has been submitted to Journal of Molecular Biology.

To meet the format of CAGI challenge, we marked number 1 as Prediction result if the effect is predicted as “no change”, marked 1-possibility\_score for “decrease” ones, marked 1+possibility\_score for “increase” ones. We cannot predict wild type or mutation containing X, so these are marked as “\*”.But for the “Standard\_Deviation” column, we have no idea how to define the value for each mutation, just leave “\*” for all mutations. We think each protein can has one SD value for all mutations’ prediction result, which is 0.608 for PTEN, and 0.63 for TPMT.